統計学入門

横田 壽

目 次

第 1 章　資料の整理 **iii**
1.1 度数分布表 (Frequency Tabulations) iii
1.2 標本の散布度，相関関係 (Dispersion, Correlation viii
1.3 相関表，回帰直線 (Correlation diagram, Regression line) xii

第 2 章　確率分布 **xvii**
2.1 確率分布 (Probability distribution) xvii
2.2 離散型の場合 (Discrete case) xviii
2.3 連続型の場合 (Continuous case) xix

第 3 章　統計的推定法 **xxiii**
3.1 統計量と標本分布 (Satatistics and sample distribution) xxiii
3.2 最尤推定 (maximum likelihood estimation) xxvi
3.3 信頼区間 (Confidence intervals) xxix
3.4 母比率の区間推定 (大標本の場合)(Confidence interval for population proportion) .xxxiii
3.5 母比率の区間推定 (小標本の場合)(Confidence interval for population proportion) . xxxv
3.6 重要な標本分布 (Important sample deistributions) xxxvii
3.7 χ^2 分布 (Chi-square distribution) xxxviii
3.8 t 分布 (Student t distribution) xl
3.9 F 分布 (F distribution) xli

第 4 章　統計的検定 **xliii**
4.1 統計的検定の考え方 (Statistical hypothesis testing) xliii
4.2 母集団が正規分布で 2 標本の場合 (Two samples from a population with normal distribution) xlvii
4.3 比率の検定 (Test of proportions) li
4.4 適合度検定 (Goodness of fit) liv
4.5 検定に用いる統計量 (Test statistic) lx

第 5 章　演習問題解答 **lxiii**

第1章　資料の整理

1.1　度数分布表 (Frequency Tabulations)

サンプリング

データの収集過程やサイコロを投げるなどの試行を繰り返すことにより結果を得る過程を**サンプリング (sampling)** といいます．サンプリングの結果えられたものを**サンプル (sample)** または**標本**といいます．

例えば，次のようなデータを得たとしましょう．

例題 1.1

表 1.1: がん患者のヘモグロビン濃度

13.6	14.8	13.7	14.2	11.5
11.9	13.8	14.6	14.2	12.7
13.4	11.5	11.9	14.8	12.7
12.4	15.3	15.2	13.5	15.0
12.4	12.0	13.8	11.7	10.0
13.2	15.5	14.0	13.5	15.0
12.7	12.9	13.7	15.1	13.5
15.7	12.7	15.7	10.9	14.0
14.8	14.0	13.8	12.7	11.9
12.0	11.4	11.1	13.7	13.2

このデータを見ただけでは，どんな結果がでたのか分かりにくいので，これらのデータを整理して分かりやすい表にすることを考えます．データの整理の方法として**度数分布表 (frequency table)** を用いることがよくありますので，度数分布表の作り方を学びます．

データの値を x_i で表すとき，x_i が現れる回数を**度数 (frequency)** といい，f_i で表すと，

$$f_1 + f_2 + \cdots + f_k = n$$

ただし，n はデータの数です．これより，度数を表にしたものを作成することができます．

可能な値	度数
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k
合計	n

しかし，データの多くは小数点以下を切り捨てたり，四捨五入したりして得たものなので，x_i という値で表を作成する代わりに，a 以上 b 未満での度数という形で表を作成します．このとき，データを a 以上 b 未満というようないくつかの区間に分けて集計するときの各区間を**階級 (class interval)** といい，$a \sim b$ で表します．そして， 区間の幅つまり $b - a$ を**階級幅 (class interval width)** といいます．それぞれの区間の端点の相加平均 $\dfrac{a+b}{2}$ を**階級値 (midpoint)** といいます．また，全標本の個数 n に対する各階級の度数の割合 f_i/n を**相対度数 (relative frequency)** いいます．さらに，統計解析のために f_i 以下の度数の合計

$$F_i = f_1 + f_2 + \cdots + f_i$$

を考えます．これを**累積度数**といいます．これらを用いて表したものが**度数分布表 (frequency distribution)** です．では，上記のデータを用いて度数分布表を作成してみましょう．データの値が 10.0〜15.7 なので，階級幅を 0.9 にとると 7 個の階級を用いることにより全てのデータを含むことが可能です．ただし，データの値が階級の境界値となるのはおかしいので，最初の階級を 9.95 から始めます．

表 1.2: 度数分布表

階級	階級値	度数	相対度数	累積度数	累積相対度数
9.95 ∼ 10.85	10.4	1	0.02	1	0.02
10.85 ∼ 11.75	11.3	6	0.12	7	0.14
11.75 ∼ 12.65	12.2	7	0.14	14	0.28
12.65 ∼ 13.55	13.1	12	0.24	26	0.52
13.55 ∼ 14.45	14.0	12	0.24	38	0.76
14.45 ∼ 15.35	14.9	9	0.18	47	0.94
15.35 ∼ 16.25	15.8	3	0.06	50	1.00

度数分布表を図（棒グラフ）で表したものを**ヒストグラム (histogram)** といいます．また，変量の小さい階級から順に度数を加えていったものを**累積度数 (cumulative distribution function)** といいます．

図 1.1: ヒストグラム

図 1.2: 累積度数分布図

Sturges の式

データ数 n に対して階級数を決める一つの目安にスタージスの式があります.

$$\text{階級数} = 1 + \frac{\log_{10} n}{\log_{10} 2}$$

この式を用いて例題 (1.1) の階級数を求めてみると，標本数 n が 50 より，階級数 k は

$$k = 1 + \frac{\log_{10} 50}{\log_{10} 2} = 1 + 3.32 \log_{10} 50 = 1 + 3.32(1.699) = 6.64 \approx 7$$

となります．また，階級幅は (最大値-最小値)/階級数で求まるので，階級幅は (15.7-10.0)/6.64 = 0.86 となります．したがって，階級幅を 0.9 とすると階級数は 7 となります.

標本の代表値

度数分布表が得られると，データ全体を視覚的に把握することができるようになります．しかしながら，それはあくまで直感的なことです．そこで，直感に頼るのではなく理論的にデータを処理するために，データの特徴を数値で表します.

代表値：分布の特徴を代表する数値

変量 x に関する n 個のデータ $x_1, x_2, \ldots, x_n$ が与えられたとき,

$$T = x_1 + x_2 + \cdots x_n = \sum_{i=1}^{n} x_i$$

を**標本総計値 (total)** といいます．また,

$$\overline{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{T}{n}$$

を**標本平均値 (mean)** といいます．また変量 x の値が $x_1, x_2, \ldots, x_k$ で，その度数が $f_1, f_2, \ldots, f_k$ で与えられているとき，標本総計値は

$$T = x_1 f_1 + x_2 f_2 + \cdots + x_n f_n$$

となるので，標本平均値は

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i f_i$$

で与えられます.

変量の測定値を，大きさの順に並べたとき，中央の位置にくるものを，**ミディアン (median)** または**中央値**といいます．データの数 n が偶数のときは第 $\dfrac{n}{2}$ 番目と第 $\dfrac{n}{2}+1$ 番目の変量の平均が中央値．また，データの数 n が奇数のときは第 $\dfrac{n+1}{2}$ 番目の変量が中央値となります.

度数が最も大きい標本値 x_i，または階級値 m_i を**モード (mode)** または最頻値といいます.

確認問題

- 例題 (1.1) の中央値を求めよ.
- 例題 (1.1) のモードを求めよ.

統計学演習問題 1

1. 次のデータについて，スタージェスの式をもちいて度数分布表・ヒストグラム・累積度数分布図を作成しよう．また，平均値，最大値，最小値，中央値，最頻値を求めよう．

コンクリート円柱の引っ張りの強さ (g/cm^2)

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360

1.2　標本の散布度，相関関係 (Dispersion, Correlation

散布度 (dispersion)

5 点満点のテストを行なったところ次のような度数分布表を得ました.

階級	f_i	f_i/n	F_i	F_i/n
0	2	0.02	2	0.02
1	13	0.13	15	0.15
2	33	0.33	48	0.48
3	35	0.35	83	0.83
4	16	0.16	99	0.99
5	1	0.01	100	1.00

また，代表値として次の表を得ました.

標本数	T	$\bar{x}$	$\sum f_i x_i^2$	s^2	s
100	253	2.53	741	1.01	1.00

この表の s^2 と s について説明します.

散布度：データが平均のまわりに集中して分布するか，平均のまわりから散らばって分布するかの程度を表わすのが，**散布度**です.

度数分布表において，各階級数 x_i $(i = 1, 2, \ldots, k)$ に対する度数を f_i とするとき，変量 x の平均 $\overline{x}$ からの偏差の平方の平均:

$$\begin{aligned} s^2 &= \frac{(x_1 - \overline{x})^2 f_1 + (x_2 - \overline{x})^2 f_2 + \cdots + (x_k - \overline{x})^2 f_k}{n} \\ &= \frac{1}{n} \sum_{i=1}^{k} (x_i - \overline{x})^2 f_i \end{aligned}$$

を**標本分散 (variance)** といいます．また，標本分散の正の平方根:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{k} (x_i - \overline{x})^2 f_i}$$

を**標準偏差 (standard deviation)** といいます．テキストによっては，標本分散の定義が

$$\begin{aligned} s^2 &= \frac{(x_1 - \overline{x})^2 f_1 + (x_2 - \overline{x})^2 f_2 + \cdots + (x_k - \overline{x})^2 f_k}{n - 1} \\ &= \frac{1}{n - 1} \sum_{i=1}^{k} (x_i - \overline{x})^2 f_i \end{aligned}$$

となっています．この 2 つの式の違いは，前者は観測対象の全ての観測値が求められた場合に用います．後者はそれ以外のときに用います.

実際の問題では階級に分ける前にすべてのデータを打ち込むので，変量 x に関する n 個のデータ $x_1, x_2, \ldots, x_n$ が与えられたとき，分散は次の式で与えられます.

$$\begin{aligned} s^2 &= \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_k-\overline{x})^2}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \end{aligned}$$

ここで分散を簡単に計算する実用的な方法として次の簡便計算法があります.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2$$

例題 1.2

上の式を導きなさい.

解答

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2-2x_i\overline{x}+\overline{x}^2) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\overline{x}^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2\right)-2\overline{x}\overline{x}+\overline{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2 \end{aligned}$$

標準偏差は平均値のまわりのデータの散らばりの大きさを表す量ですが，標準偏差が 10 点であるといっても平均点が 30 点のときと，60 点のときでは違いがあることが分かります．この違いを表す量として，**変動係数**とよばれるものがあります．変動係数はデータの平均値 $\overline{x}$ で標準偏差 s を割った割合 $\frac{s}{\overline{x}}$ で表します．したがって，変動係数は平均値に対する相対的な散らばりの大きさを表します.

例題 1.3

あるクラスの英語の試験の平均点 $\bar{x}$ は 67 で標準偏差 s_x は 8.5．また，数学の試験の平均点 $\bar{y}$ は 53 で標準偏差 s_y は 12.6 でした．このクラスの A 君の成績は英語が 75 点で数学が 68 点でした．A 君のクラスでの成績は，英語と数学のどちらの順位が上でしょうか.

解答 2 つの異なるものを比較するには，共に同じ土俵にもってこなくてはなりません．その方法として標準化とよばれるものがあります.

$$z_i = \frac{x_i-\bar{x}}{s}$$

とおくと，$\{z_i\}$ の平均は 0 に分散 s^2 は 1 になります．そこで，英語の成績と数学の成績の標準化を行うと，

$$\begin{aligned} z_{\text{english}} &= \frac{75-67}{8.5} = 0.94 \\ z_{\text{math}} &= \frac{68-53}{12.6} = 1.19 \end{aligned}$$

となり，A 君のクラスでの成績は数学の方が英語より上であるといえます．

相関関係

2 次元データの分布の特徴は 2 つの変数の平均値と分散だけでは表わすことが困難です．そこで (x, y) の n 組のデータを

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

とすると，2 つの変数の間の関係を調べるものに**共分散 (covariance)** と**相関係数 (correration coefficient)** よばれるものがあり，次のように定義されます．

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

ここで s_x は変量 x の標準偏差，s_y は変量 y の標準偏差を表わします．

確認問題

1. 相対累積度数が $p/100$ であるような標本値 x を **p パーセント点**といいます．特に，25%点 Q_1 を**第 1 四分位数**, 50%点 Q_2 を第 2 四分位数, 75%点 Q_3 を第 3 四分位数といいます．次の 20 個のデータの第 1 四分位数を求めよ．

67	54	54	66	56	65	46	35	45	45
83	72	54	58	47	60	43	82	76	92

統計学演習問題 2

1. 次のデータについて，共分散，相関係数を求めよう.

表 1.3: 二酸化硫黄と二酸化窒素の濃度

時刻	二酸化硫黄 x	二酸化窒素 y	時刻	二酸化硫黄 x	二酸化窒素 y
1	23	43	13	38	21
2	21	28	14	51	37
3	18	17	15	109	65
4	17	16	16	90	65
5	17	16	17	78	50
6	15	10	18	75	58
7	13	5	19	34	42
8	14	5	20	33	52
9	16	8	21	29	55
10	17	13	22	31	55
11	17	11	23	25	55
12	35	28	24	25	51

1.3　相関表，回帰直線 (Correlation diagram, Regression line)

相関表

ある場所で 1 時間おきに二酸化硫黄と二酸化窒素の濃度を測定しました．このとき，二酸化硫黄と二酸化窒素の濃度の間にはどんな関係があるのか調べるために準備をします．

ある時刻での二酸化硫黄の濃度を x_i, 二酸化窒素の濃度を y_i とし，x 軸に二酸化硫黄の濃度を y 軸に二酸化窒素の濃度をとって，座標 (x_i, y_i) を持つ点を図示したものを**相関図 (correration diagram)** といいます．

x_i が増加するとき，y_i も増加する傾向があるとき，x_i と y_i は**正の相関 (positive correration)** があるといいます．これに反し，x_i が増加するとき，y_i が減少する傾向があるとき，x_i と y_i は**負の相関 (negative correration)** があるといいます．

相関表

相関図ではデータの数が非常に多い場合には，その図示が困難となる場合があります．そのような場合には，2 つの変量を同時に考えた度数分布表として表わすと便利です．このような表を**相関表**といいます．

回帰直線

二酸化硫黄と二酸化窒素の関係のように，変量 x の値 $x_1, x_2, \ldots, x_n$ に，変量 y の値 $y_1, y_2, \ldots, y_n$ がそれぞれ対応していると仮定します．このとき，平面上の n 個の点:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

に対して，y 軸方向の距離が，ある意味で最も近い直線 (**回帰直線 (linear regression)**):

$$y = ax + b$$

を求めることを考えます,

もし回帰直線が求められていれば，x_i に対する y の値 (予測値 $\hat{y}_i$) は

$$\hat{y}_i = ax_i + b$$

となります．ところが x_i に対する実際の値 (観測値) は y_i です．そこでこの 2 つの値の差 $y_i - \hat{y}_i$ を予測誤差といい d_i で表わします.

$$d_i = y_i - \hat{y}_i = y_i - ax_i - b$$

ここで，

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

を最小にするような a, b の値を定め，最適合直線を求める方法を**最小２乗法 (method of least square)** といいます．

では，どうすれば $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ を最小にする a, b を求めることができるでしょうか．ここで，a, b の値が変化することにより，$\sum_{i=1}^{n}(y_i - ax_i - b)^2$ の値が変化するので，

$$F(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

とおくことにします．ここで，微分積分学の授業で学んだ微分可能な関数 $y = f(x)$ が極値をとる点では何が起きているかを思い出すと，$y' = f'(x) = 0$ となります．このことを，a と b という変数に対して行うと，

$$\begin{aligned} \frac{\partial F}{\partial a} &= \sum_{i=1}^{n} [2(y_i - ax_i - b)(-x_i)] = -2\sum_{i=1}^{n} (y_i - ax_i - b)x_i \\ \frac{\partial F}{\partial b} &= \sum_{i=1}^{n} [2(y_i - ax_i - b)(-1)] = -2\sum_{i=1}^{n} (y_i - ax_i - b) \end{aligned}$$

この式を書き直すと**正規方程式**

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i &= 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn &= 0 \end{aligned}$$

を得ます．ここで，$\overline{x} = \frac{\sum_{i=1}^{x_i}}{n}, \overline{y} = \frac{\sum_{i=1}^{y_i}}{n}$ であることに注意すると，上の２式は次のように書きなおせます．

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - bn\overline{x} &= 0 \\ n\overline{y} - an\overline{x} - bn &= 0 \end{aligned}$$

第２式から，$\overline{y} - a\overline{x} - b = 0$ となるので，これを第１式に代入すると，

$$\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - (\overline{y} - a\overline{x})n\overline{x} = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y} = a\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

となります．ここで，両辺を n で割ると，

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\overline{y} = a\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right)$$

となり，左辺は共分散 s_{xy}，右辺は a かける x の分散 as_x^2 です．したがって，

$$s_{xy} = as_x^2$$

となり，求める a は

$$a = \frac{s_{xy}}{s_x^2}$$

となります．最後に，$\overline{y} - a\overline{x} - b = 0$ より，b を求めると，

$$b = \overline{y} - a\overline{x} = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}$$

これより，x **上の** y **の回帰直線**

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

が求まります．ここまでを整理すると，

n 個のデータ $\{(x_i, y_i)\ (i = 1, 2, \ldots, n)$ について，y の x への回帰係数は

$$a_{yx} = \frac{s_{xy}}{s_x^2} = \frac{nT_{xy} - T_xT_y}{n\sum x_i^2 - T_x^2}$$

y の x への回帰直線 l の方程式は

$$y - \overline{y} = a_{xx}(x - \overline{x})$$

重回帰分析

回帰直線を用いることができるのはデータが x,y の時だけではありません．例えば，自動車の部品メーカーでは，あるセラミック部品の寸法 y を精度よく予測する課題に取り組むことになった．y を予測するための変数として，次の 3 つの変数が考えられている．

x_1 ： 注入速度
x_2 ： 材料の粒度
x_3 ： 水分量

この問題では，y を予測するために 3 つの変数 x_1, x_2, x_3 を用いることになる．このとき，y を**目的変数**，x_1, x_2, x_3 を**説明変数**といいます．この場合，回帰直線は

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

の形をとり，この式を求めることを**重回帰分析**といいます．

統計学演習問題 3

1. 次のデータ から得られる相関表より回帰直線を求めよう.

表 1.4: 二酸化硫黄と二酸化窒素の濃度

時刻	二酸化硫黄 x	二酸化窒素 y	時刻	二酸化硫黄 x	二酸化窒素 y
1	23	43	13	38	21
2	21	28	14	51	37
3	18	17	15	109	65
4	17	16	16	90	65
5	17	16	17	78	50
6	15	10	18	75	58
7	13	5	19	34	42
8	14	5	20	33	52
9	16	8	21	29	55
10	17	13	22	31	55
11	17	11	23	25	55
12	35	28	24	25	51

第2章　確率分布

2.1　確率分布 (Probability distribution)

- **確率変数** $x_1, x_2, \ldots, x_n$ なる n 個の値をとる変数 X に対して，$X = x_i$　なる確率 p_i が与えれているとき，X を**確率変数**といいます.
- **確率分布** 確率変数 X とそれに対応する確率 $P(X = x_i)$ との対応関係を**確率分布**といいます.
- **分布関数** 確率変数 X の値がある値 x までとる確率を $F(x)$ で表し，確率変数 X の**分布関数**といいます．つまり，分布関数は $F(x) = P(X \leq x)$ で与えられます.

例題 2.1

サイコロを 6 回投げるとき，$E =$「1 の目がでる」という事象のおきる確率は $P(E) = \frac{1}{6}$ で与えられる．このとき，$X =$「事象 E が発生する回数」とおくと，X は 0 から 6 までの 7 個の値をとる変数で,

$$p_i = P(X = i) = \binom{6}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{6-i}$$

で与えられます．したがって，X は確率変数で，その確率分布は **2 項分布 (binomial distribution)** とよばれ，$X \sim B(6, \frac{1}{6})$ と表します.

次の 1〜3 を満たす試行を**ベルヌーイ試行**といいます.

1. 各試行において，その事象が発生するか否かのみを問題にする
2. 各試行は統計的に独立
3. 対象とする事象が発生する確率は，各試行を通じて一定

1 回の試行において，ある事象 X が発生する確率を p とします．n 回のベルヌーイ試行列において，ちょうど i 回事象 X が発生する確率は

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

で表され，このとき X の確率分布を 2 項分布といい，$X \sim B(n.p)$ と表します.

確率変数 X のとる値が有限個または，無限個であっても自然数で番号が付けられる場合，確率変数 X は**離散型**であるという．また，確率変数 X がある区間内の全ての実数を取り得る場合，**連続型**であるという.

2.2　離散型の場合 (Discrete case)

確率変数 X のとる値を $x_1, x_2, \ldots, x_n$ とし，各事象 $(X = x_i)$ の確率を $p_1, p_2, \ldots, p_n$ とするとき，

$$P(X = x_i) = p_i \ (i = 1, 2, \ldots, n) \quad \sum p_i = 1, (p_i \geq 0)$$

で表される．これより，X の確率分布 f は

X の値 x_i	x_1	x_2	$\cdots$	x_n
$P(X = x_i) = p_i = f(x_i)$	p_1	p_2	$\cdots$	p_n

また，確率変数 X のとる値を $x_1 < x_2 < \cdots < x_n$ とするとき，その分布関数 $F(x_r)$ は次のように求められる．

$$F(x_r) = P(X \leq x_r) = p_1 + p_2 + \cdots + p_r = \sum_{i=1}^{r} p_i$$

確率分布 f と分布関数 F は次の性質をもつ．

1. $0 \leq p_i = f(x_i) \leq 1 \ (i = 1, 2, \ldots, n)$
2. $F(x_n) = P(X \leq x_n) = p_1 + p_2 + \cdots + p_n = 1$
3. $P(a < X \leq b) = F(b) - F(a)$
4. $a < b \Longrightarrow F(a) < F(b)$

平均と分散

確率変数 X の平均 (期待値) と分散は次の式で定義されます．

$$\mu = E(X) = \sum_{i=1}^{k} x_i p_i$$

$$\sigma^2 = V(X) = E\left((X - \mu)^2\right) = E(X^2) - E(X)^2$$

例題 2.2

$E(X) = \sum_{i=1}^{k} x_i p_i, \ E(Y) = \sum_{j=1}^{l} y_j q_j$ のとき，

$$E(X + Y) = E(X) + E(Y)$$

が成り立つことを示そう．

解答 $P(X = x_i, Y = y_j)$ を p_{ij} で表すと

$$\begin{cases} \sum_{j=1}^{l} p_{ij} = p_i & \sum_{i=1}^{k} p_{ij} = q_j \\ \sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} = \sum_{i=1}^{k} p_i = \sum_{j=1}^{l} q_j = 1 \end{cases}$$

これより,

$$\begin{aligned} E(X+Y) &= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_i+y_j)p_{ij} \\ &= \sum_{i=1}^{k}(x_i\sum_{j=1}^{l}p_{ij})+\sum_{j=1}^{l}(y_j\sum_{i=1}^{k}p_{ij}) \\ &= \sum_{i=1}^{k}x_ip_i+\sum_{j=1}^{l}y_jq_j = E(X)+E(Y) \end{aligned}$$

例題 2.3

$E((X-\mu)^2) = E(X^2)-(E(X))^2$ が成り立つことを示そう.

解答

$$\begin{aligned} E((X-\mu)^2) &= E(X^2-2X\mu+\mu^2) \\ &= E(X^2)-2\mu E(X)+\mu^2E(1) \\ &= E(X^2)-2E(X)E(X)+E(X)^2 = E(X^2)-E(X)^2 \end{aligned}$$

2.3 連続型の場合 (Continuous case)

確率密度関数

連続変量の確率分布において，任意の定数 a, b $(a<b)$ に対して，確率 $P_r(a \le X \le b)$ が

$$P_r(a \le X \le b) = \int_a^b f(x)dx$$

で与えられるような連続関数 $f(x)$ が $(-\infty, \infty)$ で存在するとき，この $f(x)$ を，この確率分布の**確率密度関数 (probability density function)** といいます．また，確率密度関数は次の性質を持っています.

$$f(x) \ge 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

確率分布

確率変数 X が区間 $-\infty < X \le x$ にある確率が

$$F(x) = P_r(X \le x)$$

で定められる関数 $F(x)$ を，確率変数 X の**確率分布 (probability distribution)** といいます.

平均と分散

確率変数 X の平均 (期待値) と分散は次の式で定義されます.

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$\sigma^2 = V(X) = E\left((X-\mu)^2\right) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$$

正規分布

確率変数 X の確率密度関数が

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} EXP\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],\ -\infty < x < \infty$$

で与えられるとき，確率変数 X は正規分布に従うといい，$X \sim N(\mu, \sigma^2)$ と表わします.

$X \sim N(3, 2^2)$ を表すと，次のようになります.

このままでは，比較しにくいので，**標準化 (normalization)** を行ないます.

標準化

確率変数 X の平均 $E(X)$ を 0 に，分散 $V(X)$ を 1 に直すことを標準化といいます.

標準化の方法

$$Z = \frac{X - E(X)}{\sqrt{V(X)}}$$

とおくと

$$E(Z) = 0,\ V(Z) = 1$$

になります.

$P_r(Z \leq z)$ を求めるには，$P_r(Z \leq z) = P_r(Z \leq 0) + P_r(0 < Z \leq z)$ を求めます．$P_r(Z \leq 0)$ は標準正規分布の左半分なので，その値は 0.5 となります．$P_r(0 < Z \leq z)$ の値は標準正規分布表を用いて求めます.

図 2.1: 正規分布

例題 2.4

$X \sim N(60.9, 2.9^2)$ のとき,
(1) $P(X \leq 63.8)$ を求めよ
(2) $P(62.3 < X \leq 63.0)$ を求めよ.
解答
(1) 標準化を行うと,

$$\begin{aligned} P(X \leq 63.8) &= P(\frac{X-60.9}{2.9} \leq \frac{63.8-60.9}{2.9}) \\ &= P(Z \leq \frac{2.9}{2.9}) = P(Z \leq 1) \\ &= P(Z \leq 0) + P(0 \leq Z \leq 1) = 0.5 + 0.3413 = 0.8413 \end{aligned}$$

(2)

$$\begin{aligned} & P(\frac{62.3-60.9}{2.9} < \frac{X-60.9}{2.9} \leq \frac{63.0-60.9}{2.9}) \\ = \; & P(\frac{1.4}{2.9} < Z \leq \frac{2.1}{2.9} \\ = \; & P(0.48 < Z \leq 0.72) \\ = \; & P(0 \leq Z \leq 0.72) - P(0 \leq Z \leq 0.48) \\ = \; & 0.2642 - 0.1844 = 0.0798 \end{aligned}$$

統計学演習問題 4

1. $X \sim N(80, 6^2)$ のとき，次の確率を求めよ．

(a) $P_r(X \leq 90)$

(b) $P_r(|X - 80| \leq 12)$

2. 都市 A の夏期を除く各期の一人一日当たりの水需要量は，これまでの何年かの実績からほぼ $N(210, 21^2)$ に従うことが分かっているとする．今年の一人当たりの水需要量 (夏期を除く) が 250(l/人) 以上になる確率を求めよ．

第3章　統計的推定法

3.1　統計量と標本分布 (Satatistics and sample distribution)

日本の小学 6 年生の身長を調査するとします．このとき，対象全体についての調査を**全数調査 (complete enumeration)** といいます．しかし，全数調査は労力や経費の点から不可能なことがよくあります．そこで，全数調査に代わるものとして，対象全体から何らかの方法で一部の対象を選び出し調査を行い，それにより対象全体についての推測をする方法を**標本調査 (sampling)** といいます．このとき，調査対象となる小学 6 年生の身長の集まりを**母集団 (population)** といいます．また，調査のために選び出された 6 年生の身長の集まりを**標本 (sample)** といいます．

標本抽出

日本の小学 6 年生を Π とし，小学 6 年生の各人の身長を X とすると，母集団は (Π, X) と表せます．この母集団から取り出した n 個の要素の組 $(x_1, x_2, \ldots, x_n)$ を**大きさ n の標本**といいます．このとき，個々の x_i は X と同じ分布をする確率変数 X_i が実現した数値でなければなりません．そこで，確率変数の組 $(X_1, X_2, \ldots, X_n)$ を**大きさ n の確率標本変数 (sample random variable)** といいます．確率標本変数 $(X_1, X_2, \ldots, X_n)$ に要求される数学的条件は，各 X_i が母集団 (Π, X) の X と同じ分布をする独立な確率変数であることです．では，実際に標本を選ぶときには，どのようにしたらよいのでしょうか．それには，個々の標本が全く偶然に，つまり同じ確率で現れるように選ばれる必要があります．例えば，6 人から 1 人を選ぶには，正しいサイコロを振って決めるとか，52 人から 2 人を選ぶとき，トランプのカードに各人を対応させて，よく切ったあと 2 枚を選ぶなどがあります．このようにして，標本を選ぶことを**無作為抽出**または**ランダム抽出 (random sampling)** といいます．そして，このようにして選ばれた標本を**確率標本 (probability sample)** といいます．

この母集団から無作為に抽出された標本を

$$X_1, X_2, \ldots, X_n$$

とします．標本確率変数 $X_i (i = 1, 2, \ldots, n)$ は互いに独立に母集団分布に従います．よって，

$$E(X_i) = \mu,\ V(X_i) = \sigma^2$$

となります．ここで，標本 $X_1, X_2, \ldots, X_n$ を用いて母平均と母分散を推定することを考えます．まず，素朴に考えて，$X_1, X_2, \ldots, X_n$ を n 個のデータの集まりとして，その平均と分散を求めます．す

ると，

$$\text{標本平均} \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{標本分散} \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

を得ます．このとき，$\overline{X}$ は母平均 μ を，S^2 は母分散 σ^2 を推測するのに，適当な統計量かという疑問がでます．

例題 3.1

標本平均の分散と標準偏差を求めよう．

解答

$$\begin{aligned} V(\overline{X}) &= E(\overline{X}^2) - E(\overline{X})^2 \\ &= E(\frac{1}{n}(\sum_{i=1}^{n} X_i^2)) - \mu^2 \\ &= \frac{1}{n^2}E(X_1^2 + \cdots + X_n^2 + 2(X_1X_2 + \cdots + X_{n-1}X_n)) - \mu^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i^2) + 2\sum_{1\le i,j\le n} E(X_iX_j)\right) - \mu^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}(\sigma^2 + \mu^2) + \frac{2}{n^2}\binom{n}{2}\mu^2 - \mu^2 = \frac{\sigma^2}{n} \end{aligned}$$

したがって，標本平均の標準偏差は $\frac{\sigma}{\sqrt{n}}$

定理 3.1

[チェビシェフの定理]

確率変数 X の平均値を μ，標準偏差を σ とすると，定数 $\lambda > 1$ に対して

$$P(|X - \mu| \ge \lambda\sigma) \ge \frac{1}{\lambda}$$

または

$$P(|X - \mu| < \lambda\sigma) \ge 1 - \frac{1}{\lambda^2}$$

例題 3.2

母平均 μ, 母分散 σ^2 の母集団 (Π, X) がある．ここから抽出した大きさ n の標本平均を $\overline{X}$ とする．いま $\overline{X}$ と μ との差が標準偏差 σ の $\frac{1}{5}$ より小さい確率を 0.9 以上にしたい．n をいくらにとればよいか．

解答 題意を式で表すと

$$P(|\overline{X} - \mu| < \frac{\sigma}{5}) \ge 0.9$$

一方，$\overline{X}$ の標準偏差は $\frac{\sigma}{\sqrt{n}}$ であるから，チェビシェフの定理を用いると

$$P(|\overline{X}-\mu|<\frac{\lambda\sigma}{\sqrt{n}})\geq 1-\frac{1}{\lambda^2}$$

よって，$\frac{\lambda}{\sqrt{n}}=\frac{1}{5}$ で

$$P(|\overline{X}-\mu|<\frac{\sigma}{5})\geq 1-\frac{25}{n}\geq 0.9$$

とすればよい．これから，$n\geq 250$ とすればよいことが分かる．

統計的推定 (statistical estimation)

母集団から無作為に抽出された標本

$$X_1, X_2, X_3, \ldots, X_n$$

から，標本平均

$$\overline{X}=\frac{1}{n}[X_1+X_2+X_3+\cdots+X_n]$$

標本分散

$$S^2=\frac{1}{n}[(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2$$

といった標本の統計量の値 (統計値) を用いて，母集団の分布に含まれる母数 (母平均，母分散) の値を推定することを統計的推定といいます．

点推定 (point estimation)

点推定は母数を 1 個の数値で定めようとする方法のことです．全数調査ができれば，母集団の母数は簡単に求めることができます．しかし，大事なことは，全数調査ができないときに，標本を通して母数の情報を得ることです．

母数を θ とし，これに対し大きさ n の標本変量 $\{x_1, x_2, \ldots, x_n\}$ の統計量 $T(x_1, x_2, \ldots, x_n)$ を考えます．この関数に抽出された標本値 $(X_1, X_2, \ldots, X_n)$ を代入した値 $\hat{\theta}=T(X_1, X_2, \ldots, X_n)$ でもって，θ の値であると推定することを，θ の点推定という．

不偏推定量 (unbiased estimator)

ある推定値 $\hat{\theta}=T(X_1, X_2, \ldots, X_n)$ について，

$$E(\hat{\theta})=E(T(X_1, X_2, \ldots, X_n))=\theta$$

のとき，$\hat{\theta}$ を θ の**不偏推定量**という．

母集団 $N(\mu, \sigma^2)$ において，次の統計量は不偏推定量である．

$$\bar{X}=\frac{1}{n}\sum_{n=1}^{n}X_i,\ U^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

この意味で，U^2 を**不偏分散**という．

例題 3.3

$\bar{X}$ は不偏推定量であることを示そう.
解 $E(X+Y)=E(X)+E(Y)$ より,

$$nE(\bar{X})=E(X_1)+E(X_2)+\cdots+E(X_n)=n\mu$$

したがって, $E(\bar{X})=\mu$.

しかし, 母分散 σ^2 を推定するには, 不偏性とは異なる立場をとると, U^2 よりも標本分散

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{n-1}{n}U^2$$

の方が良いこともあります. このようになる理由は, n を増やしても U^2 や S^2 が σ^2 を中心とする狭い区間に入る確率がなかなか大きくならないことにあります.

3.2　最尤推定 (maximum likelihood estimation)

母集団分布の形が分かっているがその母数が未知であるときに, n 個の標本値 $x_1,x_2,\ldots,x_n$ を母集団分布に従う確率変数 $X_1,X_2,\ldots,X_n$ がとることは最も起こりやすい (maximum likelihood) という条件を用いてその母数を決めようとするものである.

例題 3.4

ポワソン母集団から大きさ 3 の独立な標本を無作為に抽出したとき, その値が x_1,x_2,x_3 であったとする. この標本値から母平均 μ を推定しよう.
解 標本値 x_1,x_2,x_3 は, 母集団と同じポワソン分布に従い, かつ互いに独立な確率変数 X_1,X_2,X_3 たとった値だと考えられる. そのような値をとる確率 $P(X_1=x_1,X_2=x_2,X_3=x_3)$ を L とすると, X_1,X_2,X_3 は独立より,

$$L=P(X_1=1)P(X_2=x_2)P(X_3=x_3)=e^{-\mu}\frac{\mu^{x_1}}{x_1!}\cdot e^{-\mu}\frac{\mu^{x_2}}{x_2!}\cdot e^{-\mu}\frac{\mu^{x_3}}{x_3!}=e^{-3\mu}\frac{\mu^{x_1+x_2+x_3}}{x_1!x_2!x_3!}$$

となる. ここで, この確率が最も起こりやすい μ を求める. つまり, L が最大となるような μ を求める. x_1,x_2,x_3 は標本値として既知であるから, μ の関数としての $L=L(\mu)$ は,

$$\frac{dL}{d\mu}=0$$

のときに最大となる. したがって,

$$\begin{aligned}\frac{dL}{d\mu} &= -3e^{-3\mu}\frac{\mu^{x_1+x_2+x_3}}{x_1!x_2!x_3!}+(x_1+x_2+x_3)e^{-3\mu}\frac{\mu^{x_1+x_2+x_3-1}}{x_1!x_2!x_3!}\\ &= -3L+\mu^{-1}(x_1+x_2+x_3)L=\frac{L}{\mu}(-3\mu+x_1+x_2+x_3)=0\end{aligned}$$

より,

$$\mu=\frac{1}{3}(x_1+x_2+x_3)$$

が母平均の推定値である.

このようにして得られた推定量を**最尤推定量**といい，推定値を得るために考えた関数 L を**尤度関数**といいます.

例題 3.5

$N(\mu,\sigma^2)$ に従う正規母集団から，大きさ n の独立な標本を無作為抽出したところ，その標本値が $x_1,x_2,\ldots,x_n$ であった．母分散 σ^2 が既知のときの母平均 μ の最尤推定量を求めよ.

解 $N(\mu,\sigma^2)$ の確率密度は

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

である．n 個の標本は互いに独立なので

$$L=P(X_1=x_1)\cdots P(X_n=x_n)=\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{2\sigma^2}\right\}\right)$$

ここで，$x_1,x_2,\ldots,x_n,\sigma^2$ は既知だから，

$$\begin{aligned}\frac{dL}{d\mu} &= -\frac{1}{2\sigma^2}\{2(\mu-x_1)+2(x_2-\mu)+\cdots+2(\mu-x_2)\}L\\ &= -\frac{1}{\sigma^2}\{n\mu-(x_1+x_2+\cdots+x_n)\}L=0\end{aligned}$$

したがって,

$$\mu=\frac{1}{n}(x_1+x_2+\cdots+x_n)=\bar{x}$$

が最尤推定量となる.

統計学演習問題 5

1. 次のデータの不偏分散を求めよう．また，標準偏差を求めよう．

$$110, 121, 133, 124, 126, 118, 112, 125, 131, 120(\text{cm})$$

3.3 信頼区間 (Confidence intervals)

区間推定 (interval estimation)

母数 θ がある区間 $[\theta_1, \theta_2]$ に入るだろうと推定するのが区間推定です．詳しくいうと，母数 θ を推定するために，母集団から無作為に抽出された標本から 2 つの統計値 θ_1, θ_2 を定める．このとき，あらかじめ指定された小さな確率 α $(0 < \alpha < 1)$ に対して，常に

$$P_r(\theta_1 < \theta < \theta_2) = 1 - \alpha$$

が満たされるとき，区間 (θ_1, θ_2) を θ の**信頼区間 (Confidence interval)**，θ_1, θ_2 を**信頼限界 (Confidence limit)**，$100(1-\alpha)\%$ を**信頼係数 (Confidence coefficient)** または**信頼度 (Confidence level)** といいます．信頼区間 $[\theta_1, \theta_2]$ を求めることを**区間推定 (Interval estimation)** といいます．

θ は一定値ですが，区間 $[\theta_1, \theta_2]$ は標本によっていろいろ変わり，この区間に θ が入る確率が $1-\alpha$ です．

区間推定法 (interval estimation method)

母集団が正規分布 $N(\mu, \sigma^2)$ に従い，母分散 σ^2 が分っているとします．このとき，母平均 μ はどの範囲にあるかを，どのくらい信頼できるかを考えて表わしてみましょう．

準備

標本 $X_1, X_2, \ldots, X_n$ が，$X_i \sim N(\mu, \sigma^2)$ のとき，

$$E(\overline{X}) = \mu,\ V(\overline{X}) = \frac{\sigma^2}{n}$$

より

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

と表せます．また，

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ の期待値 } E(S^2) = \frac{n-1}{n}\sigma^2$$

より

$$S'^2 = \frac{n}{n-1}S^2 \text{ の期待値 } E(S'^2) = \sigma^2$$

と表せます．

母平均 μ の区間推定 (σ^2 既知)

ここでは $\alpha = 0.05$ つまり，95%信頼区間を推定します．まず，

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

より標準化を行なうと，

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

これより，

$$P_r\left(|Z| \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha = 0.95$$

ここで，$z_{\frac{\alpha}{2}}$ は，

$$P_r(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

を満たす点です．このとき，$z_{\frac{\alpha}{2}}$ を標準正規分布表の両側確率で求めると，$\alpha = 0.05$ のとき，$z_{\frac{\alpha}{2}}$ は

$$z_{\frac{\alpha}{2}} = 1.96$$

となります．よって求める信頼区間は次の不等式を満たします．

$$|Z| = |\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}| \leq z_{\frac{\alpha}{2}}$$

この不等式を μ について解くと

$$\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}$$

を得ます．これが母平均 μ の信頼区間となります．

図 3.1: 正規分布

例題 3.6

標本 $28, 24, 31, 27, 22$ が与えられたとして，標準偏差が 2.5 である正規母集団の平均に対する 95% 信頼区間を求めよう．

解答 標準偏差が 2.5 より，母分散 $\sigma^2 = 6.25$ は既知である．この母集団から無作為に選んだ標本 X_i は $X_i \sim N(\mu, 6.25)$ の正規分布に従っていると考えることができる．したがって，

$$\bar{X} \sim N(\mu, \sigma^2/5)$$

となる．ここで，$\bar{X}$ を求めると，

$$\bar{X} = \frac{1}{4}[145.3 + 145.1 + 145.4 + 146.2] = \frac{132}{4} = 145.25$$

標準化を行なうと，

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/4}} \sim N(0, 1)$$

となる．95%信頼区間より，$P_r(|Z| \leq z_{\frac{\alpha}{2}}) = 0.95$. また，$z_{\frac{0.05}{2}} = 1.96$. したがって，

$$\bar{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{5}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{5}}$$

$$26.4 - 1.96\sqrt{6.25/5} \leq \mu \leq 26.4 + 1.96\sqrt{6.25/5}$$

$$24.21 \leq \mu \leq 28.59$$

次に，母集団が正規分布に従うことは分かっているが母分散 σ^2 が不明である場合を考えます.

平均値の区間推定 (σ^2 未知)

ここでは $\alpha = 0.05$ つまり，95%信頼区間を推定します．この場合，2 つの母数 μ, σ^2 が必要となりますが，σ^2 が未知なので，σ^2 を推定する不偏分散 S'^2 を σ^2 の代わりに用います．すると，母分散に無関係に

$$T = \frac{\bar{X} - \mu}{\sqrt{S'^2/n}}$$

は，自由度 $n-1$ の t 分布に従うことが知られています．これより，

$$P_r(|T| \leq t_{n-1,\alpha/2})) = 1 - \alpha$$

となります．ここで，$t_{n-1,\alpha/2}$ は，

$$P_r(T \geq t_{n-1,\alpha/2}) = \frac{\alpha}{2}$$

を満たす点である．このとき，$t_{n-1,\alpha/2}$ を t 分布表の両側確率で求めると，$\alpha = 0.05$，$n = 10$ のとき，$t_{9,0.05/2}$ は

$$t_{9,0.05/2} = 2.26$$

よって求める信頼区間は次の不等式を満たします.

$$|\frac{\bar{X} - \mu}{\sqrt{S'^2/n}}| \leq t_{n-1,\alpha/2}$$

この不等式を μ について解くと

$$\overline{X} - t_{n-1,\alpha/2}\sqrt{\frac{S'^2}{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2}\sqrt{\frac{S'^2}{n}}$$

を得ます.

統計学演習問題 6

1　ある水域の一定区間における水質 BOD(ppm) はほぼ正規分布に従い，その母分散は $\sigma^2 = 6.25(ppm)^2$ であることがわかっている．いま $n = 15$ 個の標本をとり，標本平均 $\bar{X} = 7.2ppm$ を得た．このとき信頼度 95% で，この水質の母平均の区間推定をせよ．

2　標本 $145.3, 145.1, 145.4, 146.2$ が与えられたとして，母平均が 146 で標準偏差が 2.5 ある正規母集団の平均に対する 95% 信頼区間を求めよう．

3　標本 $145.3, 145.1, 145.4, 146.2$ が与えられたとして，母平均が 146 である正規母集団の平均に対する 95% 信頼区間を求めよう．

3.4 母比率の区間推定 (大標本の場合)(Confidence interval for population proportion)

母集団の中で，ある属性に対して事象 A の起こる割合 p を事象 A の**母比率**といいます．母比率が p の二項母集団から抽出された大きさ n の標本を $(X_1, \ldots, X_n)$ とします．ここで，

$$X_i = \begin{cases} 1 & A \text{ のとき} \\ 0 & \bar{A} \text{のとき} \end{cases}$$

とします．このとき，$X = X_1 + \cdots + X_n$ とすると，X は標本中 A であるものの個数を表す統計量で，$\dfrac{X}{n}$ は事象 A の**標本比率**といいます．

$\frac{X}{n}$ は母比率 p の不偏推定量である

母比率 p の二項母集団から大きさ n の標本 $(X_1, \ldots, X_n)$ をとり，$X = X_1 + \cdots + X_n$ とすると X は二項分布 $B(n,p)$ に従います．ここで n が十分大きいときにはラプラスの定理によって，X は近似的に正規分布 $N(np, np(1-p))$ に従い，標本比率 $\frac{X}{n} = \hat{p}$ は近似的に正規分布 $N\left(p, \frac{p(1-p)}{n}\right)$ に従います．よって，標準化を行うと

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

したがって，

$$P\left(\left|\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

が成り立ちます．この式を書き直すと

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq p + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

この両辺は母数 p を含んでいるが，n が非常に大きいときには $\hat{p}$ で近似できるので，母比率 p の信頼度 $100(1-\alpha)\%$ の信頼区間は

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

となります．

定理 3.2

[ラプラスの定理] $X \sim B(n,p)$ のとき，十分大きな n に対して

X は近似的に $N(np, np(1-p))$ に従う．

例題 3.7

サイコロを 600 回投げたところ，1 の目が 108 回出たという．1 の目が出る母比率 p を信頼度 95% で区間推定せよ．

解答 標本比率は $\hat{p} = \frac{108}{600} = 0.18$．また，$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96$ であるから，上記の公式に代入すると

$$\left(0.18 - 1.96\sqrt{\frac{(0.18)(0.82)}{600}},\ 0.18 - 1.96\sqrt{\frac{(0.18)(0.82)}{600}}\right)$$

より，$(0.15, 0.21)$ となります．

3.5 母比率の区間推定 (小標本の場合)(Confidence interval for population proportion)

標本の大きさが大きければ大きいほど狭い区間で母比率の推定が可能となる．したがって，多くの標本があればよいのですが，薬品のような人体実験の場合，標本を多くとることが難しい場合があります．標本の大きさ n が小さいとラプラスの定理を用いることができません．そこで，標本の大きさ n が小さい場合には，次のような方法をとって区間推定を行います．母比率 $p = P(A)$ の二項母集団から大きさ n の標本を抽出したとき，A であるものの個数 (標本和) が x であったとします．n があまり大きくないとき，母比率 p の信頼度 $100(1-\alpha)\%$ の信頼区間 (p_1, p_2) は次のようにして得ることができます．

$$\left(\frac{n_2}{n_1 f_1 + n_2}, \frac{m_1 f_2}{m_1 f_2 + m_2} \right)$$

ただし，$n_1 = 2(n - x + 1)$, $n_2 = 2x$ とし，自由度 (n_1, n_2) の F 分布に従う確率変数 F が

$$P(F > f_1) = \frac{\alpha}{2}$$

となる f_1 を F 分布表から求める．

統計学演習問題 7

1　あるテレビ番組が，無作為標本 900 台のテレビのうち 180 台で見られていることが分かった．この番組の視聴率を 95%の信頼度で区間推定せよ．

2　ある大学の学生から無作為に 300 人を選んで，アルバイトをしているかどうかの調査をしたところ，187 人が何かのアルバイトをしていた．その大学の全学生のうちアルバイトをしている学生の比率の信頼度 95%の信頼区間を求めよ．

3.6 重要な標本分布 (Important sample deistributions)

第 2 章で，確率分布の基礎となる 2 項分布と正規分布の話をしました．ここでは，まず，正規分布の性質について考えます．

正規分布の加法性

確率変数 X, Y が独立で，それぞれ正規分布 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ に従うとき，和 $aX + bY$ は正規分布

$$N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

に従う．

定理 3.3

$X_1, X_2, \ldots, X_n$ は互いに独立で正規分布 $N(\mu, \sigma^2)$ に従えば，

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

証明 正規分布の加法性より，

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$$

したがって，

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N(\mu, \frac{n\sigma^2}{n^2}) = N(\mu, \frac{\sigma^2}{n})$$

例題 3.8

ある年齢の生徒の身長は 120cm を平均値として，標準偏差が 4.5cm の正規分布に従っているとする．このとき，50 人の身長の平均が 120.6cm よりも大きくなる確率を求めよ．**解** X_i をある年齢の生徒の身長とすると，$X_i \sim N(120, 4.5^2)$．50 人の相加平均 $\overline{X}$ は定理 3.6 より $\overline{X} \sim N(120, \frac{4.5^2}{50})$．よって

$$\begin{aligned} P(\overline{X} > 120.5) &= P(Z > \frac{120.5 - 120}{\frac{4.5}{\sqrt{50}}}) = P(Z > 0.9428) \\ &= 0.5 - P(0 < Z < 0.9528) = 0.5 - 0.32710 \approx 0.173 \end{aligned}$$

次に，X_i が正規分布に従わない場合を考える．

[**中心極限定理**] $X_1, X_2, \ldots, X_n$ が互いに独立で，平均値 μ，分散 σ^2 の同じ確率分布に従うとする．このとき，n が十分大きいならば

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \text{ 近似的に } N(\mu, \frac{\sigma^2}{n}) \text{ に従う}$$

$n > 100$ ならば近似度はよくなる．

3.7　χ^2 分布 (Chi-square distribution)

χ^2 分布は 1 つの自然数 n を含む連続型分布で，$\chi^2(n)$ と表し n をその自由度という。χ^2 分布の密度関数 $f_n(x)$ は次の式で与えられる。

$$f_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

ここで，**ガンマ関数** $\Gamma(x)$ は

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \ (x > 0)$$

で定義される。

χ^2 分布の名前は次の性質から来ている。

定理 3.4

確率変数 $X_1, X_2, \ldots, X_n$ が同一の標準正規分布 $N(0,1)$ に従い，互いに独立ならば，その統計量

$$\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

は自由度 n の χ^2 分布に従う。その期待値と分散は

$$E(\chi_n^2) = n, \ V(\chi_n^2) = 2n$$

定理 3.5

[χ^2 分布の加法性] χ_n^2, χ_m^2 がそれぞれ自由度 n,m の χ^2 分布に従い，互いに独立ならば，$\chi^2 = \chi_n^2 + \chi_m^2$ は自由度 $n+m$ の χ^2 分布に従う。

標本分散 S^2 に関して，次の定理がある。

定理 3.6

$N(\mu, \sigma^2)$ の正規分布に従う母集団から無作為で得た標本を $\{X_1, X_2, \ldots, X_n\}$ とすると，

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{nS^2}{\sigma^2}$$

は自由度が $n-1$ の χ^2 分布に従って分布する。

例題 3.9

母集団が正規分布であるとする．n が 20 の標本から標本分散を求めたところ，その値は 1.5 であった。母分散が 1 のとき，標本分散が 1.5 より大きい確率を求めよ．

解 $P(S^2 \geq 1.5)$ を求める。$X_i \sim N(\mu, 1)$ より，

$$Y = \frac{20S^2}{1} = \sum_{i=1}^{20}(X_i - \bar{X})^2$$

は自由度 19 の χ^2 分布に従う。したがって，

$$\begin{aligned} P(S^2 \geq 1.5) &= P(20S^2 \geq 28.5) \\ &= P(\chi^2_{19} \geq 28.5) \end{aligned}$$

ここで，χ^2 分布表を用いると，$P(\chi^2_{19} > 27.20) = 0.10$ で $P(\chi^2_{19} > 30.14) = 0.05$ より，

$$P(\chi^2_{19} \geq 28.5) = 0.05 + \frac{28.5 - 27.20}{30.14 - 27.20}(0.10 - 0.05) \approx 0.07$$

3.8　t 分布 (Student t distribution)

定義 3.1

確率密度関数 $f_n(x)$ が

$$f_n(x) = \frac{\gamma(\frac{n+1}{2})}{\sqrt{n\pi}\gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}} \ (n \geq 1)$$

で与えられる分布 T_n を**自由度 n の t 分布**という．$n \geq 3$ のとき，その期待値と分散は

$$E(T_n) = 0,\ V(T_n) = \frac{n}{n-2}$$

となる．

定理 3.7

$X_1, X_2, \ldots, X_n$ がいずれも正規分布 $N(\mu, \sigma^2)$ に従う互いに独立な確率変数とする．このとき，

$$U^2 = S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

とし，正の平方根を U とすると，

$$T = \frac{\overline{X} - \mu}{U/\sqrt{n}}$$

は自由度 $n-1$ の t 分布に従う。

定理 3.8

Z を標準正規確率変数，χ_n^2 を自由度 n の χ^2 確率変数とする。さらに，Z と χ_n^2 が互いに独立ならば，標本分布

$$T_n = \frac{Z}{\sqrt{\chi_m^2/n}}$$

は自由度 n の t 分布に従う。

自由度 n の t 分布を $t(n)$ と表す.

t 分布の正規近似

t 分布は自由度が大きければ標準正規分布で近似でき，

$$P_r(T_n \leq c) \approx P_r(Z \leq c)$$

となる。χ_n^2 は n 個の独立な確率変数の和であるから，n が大きければ χ_n^2/n は大数の法則により，1 に収束する。T_n 確率変数の分母が 1 に近づくから，T_n 確率変数は分子の標準正規確率変数と変わらなくなる。

3.9 F 分布 (F distribution)

定義 3.2

確率密度関数 $f_{m,n}(x)$ が

$$f_{m,n}(x) = \begin{cases} \frac{\gamma(\frac{m+n}{2})}{\gamma(\frac{m}{2})\gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} & (x>0) \\ 0, \end{cases}$$

で与えられる分布 $F_{m,n}$ を**自由度対** (m,n) **の** F **分布**という。その期待値と分散は

$$\begin{aligned} E(F_{m,n}) &= \frac{m}{n-2} \ (n>2) \\ V(F_{m,n}) &= \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \ (n>4) \end{aligned}$$

となる。

定理 3.9

$X_{11}, X_{12}, \ldots, X_{1n_1}$ がいずれも正規分布 $N(\mu_1, \sigma_1^2)$ に従う互いに独立な n_1 個の確率変数で，その相加平均を $\overline{X}_1$，不偏分散を U_1^2 とする．これらと独立に，$X_{21}, \ldots, X_{2n_2}$ は正規分布 $N(\mu_2, \sigma_2^2)$ に従う互いに独立な n_2 個の確率変数を考え，その相加平均を $\overline{X}_2$，不偏分散を U_2^2 とする．

$$\begin{aligned} \overline{X}_1 &= \frac{1}{n_1 \sum_i X_{1i}}, \ U_1^2 = \frac{1}{n_1 - 1} \sum_i (X_{1i} - \overline{X}_1)^2 \\ \overline{X}_2 &= \frac{1}{n_2 \sum_i X_{2i}}, \ U_2^2 = \frac{1}{n_2 - 1} \sum_i (X_{2i} - \overline{X}_2)^2 \end{aligned}$$

このとき，

$$F = \frac{U_1^2/\sigma_1^2}{U_2^2/\sigma_2^2} = \frac{\sigma_2^2 U_1^2}{\sigma_1^2 U_2^2} \sim F(n_1 - 1, n_2 - 1)$$

例題 3.10

$F_{n_2}^{n_1}(0.05) = F_{10}^{5}(0.05)$ を求めよ.

解

$$F_{10}^{5}(0.05) = 3.3258$$

$F_{n_2}^{n_1}(1-\alpha)$ を求めるには，次の公式を用いる

$$F_{n_2}^{n_1}(1-\alpha) = \frac{1}{F_{n_1}^{n_2}(\alpha)}$$

例題 3.11

$F_{11}^{5}(1-0.05)$ を求めよ.

解

$$F_{11}^{5}(1-0.05) = \frac{1}{F_{5}^{11}(0.05)} = \frac{1}{3.10} = 0.32$$

第4章　統計的検定

4.1　統計的検定の考え方 (Statistical hypothesis testing)

超心理学では透視実験に ESP カードを用います．カードは 5 種類からなっています．そこで 1 枚のカードを引いて裏向きに置いて，このカードの種類をあてさせます．カードを元に戻し同じ実験を 5 回繰り返したところ，ある学生は 3 回的中しました．そこで問題です．この学生の透視能力についてどのような判断を下すべきか考えてみましょう．

2 つの結論が考えられます．

結論 1. 透視能力が無くても，5 回中 3 回ぐらいは偶然でも的中すると考えられるので，これだけのデータでは透視能力があるとはいえない．

結論 2. 5 回中 3 回的中することは滅多にないことだから透視能力があると考える方がもっともらしい．

この 2 つの結論のどちらを選ぶべきかの基準に確率が用いられます．

まずこの学生がカードをあてる確率は毎回一定で，その確率は p とします．5 回の実験中的中する回数を X とすると，$X \sim B(5,p)$ に従います．ここで透視能力がないということは $\lceil p = 0.2 \rfloor$，透視能力があるということは $\lceil p > 0.2 \rfloor$ と表わせます．

そこで

$$H_0 : p = 0.2 \quad (\text{透視能力がない})$$

と仮定してみます．

ここで 5 回中 3 回以上的中する確率は

$$\begin{aligned} P_r(X \geq 3) &= P_r(X=3) + P_r(X=4) + P_r(X=5) \\ &= \binom{5}{3}(0.2)^3(0.8)^2 + \binom{5}{4}(0.2)^4(0.8) + \binom{5}{5}(0.2)^5 = 0.057922 \end{aligned}$$

となり，この確率を**有意確率 (significant probability)** といいます．

実験の結果がそれほど稀な現象ではない，つまり有意確率がそれほど小さくないと判断した場合は結論 1 を得ます．このことを仮説 H_0 を**容認する**といいます．

実験の結果がきわめて稀な現象である，つまり 有意確率がきわめて小さいと判断した場合は結論 2 になります．このことを仮説 H_0 を**棄却する**といいます．

有意確率がどの程度小さければ，H_0 を棄却したらよいかという基準を**有意水準 (significance level)** α とよび，α として $0.05, 0.01$ 等が良く用いられます．この問題で有意水準を 0.05 とすると，仮説 H_0 は棄却されない (容認されます)．つまりこの学生は透視能力がないと判断されます．

仮説 H_0 が棄却される X の範囲は $P_r(X \geq 4) = 0.00672$ より $X \geq 4$ です．この範囲を**棄却域 (critical region)** といいます．また $H_0 : p = 0.2$ を**帰無仮説 (null hypothesis)**，$H_1 : p > 0.2$ を**対立仮説 (alternative hypothesis)** といいます．

母数 θ に関する帰無仮説 $H_0 : \theta = \theta_0$ に対し対立仮説として次の 3 つがあります．

$$H_1 : \theta > \theta_0,\ H_1 : \theta < \theta_0,\ H_1 : \theta \neq \theta_0$$

母数 θ に関する検定の手順

1. 帰無仮説，対立仮説を立てる．
2. 有意水準 α を定める．
3. 帰無仮説のもとで検定に用いる統計量の分布を求める．
4. 棄却域を定める．
5. 検定統計量の実現値が棄却域に入るときは帰無仮説を棄却する．

(1) 母平均の検定

母集団が正規分布 $N(\mu, \sigma^2)$ に従う，正規母集団から抽出された大きさ n の標本変量 $(X_1, X_2, \ldots, X_n)$ を考えます．このとき，これらの相加平均 $\bar{X}$ は $N(\mu, \frac{\sigma^2}{n})$ に従がいます．ここで，母平均 μ についての検定を考えます．

(a) σ^2 が既知の場合の μ の検定 (有意水準 α)

この場合は次の標本分布を用います．

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

(b) σ^2 が未知の場合の μ の検定 (有意水準 α)

この場合は次の標本分布を用います．

$$T = \frac{\bar{X} - \mu}{\sqrt{S'^2/n}} \sim t(n-1)$$

(2) 母分散の検定

母集団が正規分布 $N(\mu, \sigma^2)$ に従う，正規母集団から抽出された大きさ n の標本変量 $(X_1, X_2, \ldots, X_n)$ を考えます．ここで，母分散 σ^2 についての検定を考えます．

(a) μ が既知の場合の σ^2 の検定 (有意水準 α)

この場合は次の標本分布を用います．

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_{\alpha,n}$$

(b) μ が未知の場合の σ^2 の検定 (有意水準 α)

この場合は次の標本分布を用います．

$$\chi^2 = \frac{nS^2}{\sigma^2} \sim \chi^2_{\alpha,n-1}$$

noindent **棄却域**

(a) σ^2 が既知の場合の μ の検定 (有意水準 α)

帰無仮説 $H_0 : \mu = \mu_0$ に対し次の 3 通りの対立仮説を考えます.

$$(1)\ H_1 : \mu > \mu_0\ (2)\ H_1 : \mu < \mu_0\ (3)\ H_1\ \mu \neq \mu_0$$

帰無仮説のもとで (つまり，$\mu = \mu_0$ のとき)

$$Z_0 = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

となります．有意水準 α に対して,

$P_r(Z_0 > z_\alpha) = \alpha$ であるような z_α を正規分布表から読み取れば，有意水準 α の対立仮説 (1) に対する帰無仮説の棄却域は $Z_0 > z_\alpha$ となります.

これより各対立仮説に対する帰無仮説の棄却域は次の通りです.

(1) の場合　$Z_0 > z_\alpha$

(2) の場合　$Z_0 < z_\alpha$

(3) の場合　$|Z_0| > z_{\frac{\alpha}{2}}$

例題 4.1

ある大学では一年生に対して毎年同じテストを行なっている．昨年度の一年生の成績は平均 64.5, 分散 20 の正規分布に従っている．今年度の一年生にも同じテストを行ない，無作為に 8 人抽出したところ点数は次の通りであった.

66	73	55	69	70	67	62	71

今年度の一年生の平均点は昨年度より高いか有意水準 5% で検定せよ．ただし，今年度の分散は，昨年度と変わらないものとする.

解

今年度の 1 年生の平均点を μ とおき，昨年度より高いかの検定を行なうので，以下の帰無仮説と対立仮説を立てる.

$$\begin{aligned} H_0 &: \mu = 64.5 \\ H_1 &: \mu > 64.5 \end{aligned}$$

今年度の標本 X_i は $X_i \sim N(\mu, 20)$ と考えることができる．したがって,

$$\bar{X} = \frac{1}{8}[66 + 73 + 55 + 69 + 70 + 67 + 62 + 71] = 533/8 = 66.625$$

また，$\bar{X} \sim N(\mu, 20/8)$. 帰無仮説のもとで，標準化を行なうと,

$$Z_0 = \frac{66.625 - 64.6}{\sqrt{20/8}} = \frac{2.025}{1.58} = 1.28$$

図 4.1: 正規分布

これより，Z_0 は帰無仮説の棄却域に入っていない．したがって，帰無仮説を棄却できない．言い換えると，今年度の 1 年生は昨年度よりも優秀であるとは，結論付けることができない．

統計学演習問題 8

1 ある工場での経験によると 7mm ボルトの規格の製品の寸法はほぼ正規分布をしており，その標準偏差は 0.20mm であるという．ある日の製品から 16 個のボルトを無作為抽出したところ，その寸法の平均が 7.09mm であった．この日の製品の寸法の平均は規格から外れているか．有意水準 $\alpha = 0.05$ で検定せよ．またこの日の製品の寸法の平均 μ の 95% 信頼区間を求めよう．

2 1 台の機械が製造する鋼球の直径 (単位 mm) は正規分布に従っており，その分散は従来の経験から 0.0016 であるといわれている．ある日この機械が製造した鋼球から 8 個を抽出しその直径を測定したところ次のようになった．

11.97 12.02 12.06 12.03 11.99 11.98 12.12 12.05

この日の製品の直径の分散は 0.0016 より大きいといえるか．有意水準 $\alpha = 0.05$ で検定せよ．またこの日の製品の直径の分散の 95% 信頼区間を求めよう．

4.2 母集団が正規分布で 2 標本の場合 (Two samples from a population with normal distribution)

正規母集団 X は母平均 μ_1，母分散 σ_1^2 であるとします．正規母集団から無作為抽出した標本を $X_1, X_2, \ldots, X_{n_1}$，標本平均を $\bar{X}$，標本分散を S_1^2 とする．同様に正規母集団 Y は 母平均 μ_2，母分散 σ_2^2 であるとします．正規母集団 Y から無作為抽出した標本を $Y_1, Y_2, \ldots, Y_{n_2}$，標本平均を $\bar{Y}$，標本分散を S_2^2 とすると，

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}),\ \bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

となります．

1. **母平均の差 $\mu_1 - \mu_2$ の検定**

(a) σ_1^2, σ_2^2 既知の場合

$\bar{X} - \mathbf{Y}$ は正規分布の加法性より，正規分布 $N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ に従がいます．よって，

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

例題 4.2

A 校から 30 人，B 校から 50 人の標本を抽出して身長を調べたところ，それぞれ平均値 148.2cm, 146.4cm であった．この年齢の生徒の身長は，標準偏差 4.8cm の正規分布に従って分布するとする．両校の身長に有意差があるか，有意水準 0.05 で検定せよ．

解 A 校の生徒は $N(\mu_1, 4.8^2)$, B 校の生徒は $N(\mu_2, 4.8^2)$ で，大きさ $n_1 = 30, n_2 = 50$ である．したがって，

$$\bar{X} \sim N(\mu_1, \frac{4.8^2}{30}), \bar{Y} \sim N(\mu_2, \frac{4.8^2}{50})$$

検定するないようは，題意より両校の身長に有意差があるかということより，

1 $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

2 有意水準 $\alpha = 0.05$

3 統計量

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

4 H_0 のもとで,

$$Z_0 = \frac{148.2 - 146.4}{\sqrt{4.8^2/30 + 4.8^2/50}} = 1.6238$$

5 $z_{0.05/2} = 1.96$ より,

$$Z_0 = 1.62 < z_{0.05/2} = 1.96$$

したがって，H_0 は棄却されない.

(b) σ_1^2, σ_2^2 が未知だが $\sigma_1^2 = \sigma_2^2$ とみなせる場合.

$X_1, \ldots, X_{n_1},\ Y_1, \ldots, Y_{n_2}$ に対して不偏分散をそれぞれ $S_1'^2,\ S_2'^2$ とします.

$$S_1'^2 = \frac{1}{n_1 - 1}\sum_i (X_i - \bar{X})^2,\ S_2'^2 = \frac{1}{n_2 - 1}\sum_i (Y_i - \bar{Y})^2$$

これを合併した不偏分散

$$\hat{\sigma}'^2 = \frac{(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2}{n_1 + n_2 - 2}$$

を考えます．これは，両方の標本分散

$$S_1^2 = \frac{1}{n_1}\sum_i (X_i - \bar{X})^2,\ S_2^2 = \frac{1}{n_2}\sum_i (Y_i - \bar{Y})^2$$

を用いて

$$\hat{\sigma}^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

としても同じです．この $\hat{\sigma}^2$ を (a) で用いた式に代入すると,

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2/n_1 + \hat{\sigma}^2/n_2}} \sim t(n_1 + n_2 - 2)$$

これは，自由度 $n_1 + n_2 - 2$ の t 分布に従がうことが分かっています．これを利用して，母平均の差の検定を行うことができます.

例題 4.3

A,B.2 つの方法で化学物質を作ろうとした．それぞれ 5 回ずつ実験したらその純度の平均値および分散は次の通りであった.
平均値 $\bar{X}_A = 97.5\%,\ \bar{X}_B = 95.3\%$
分散　$S_A^2 = 1.23\%^2, S_B^2 = 1.56\%^2$
製法によって純度に差があるだろうか．有意水準 0.05 で検定せよ.

4.2. 母集団が正規分布で 2 標本の場合 (Two samples from a population with normal distribution)xlix

解 それぞれの母純度を $\mu_A\%, \mu_B\%$ とする．これらの分析値は，同じ分散の正規分布にしたがって分布するとする．

1 $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

2 有意水準 $\alpha = 0.05$

3 統計量

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2/n_1 + \hat{\sigma}^2/n_2}} \sim t(n_1 + n_2 - 2),\ \hat{\sigma}^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}.$$

4 H_0 のもとで，$\hat{\sigma}^2 = \frac{5(1.23)+5(1.56)}{5+5-2} = 1.7428\%$

$$T_0 = \frac{97.5 - 95.3}{\sqrt{1.7438/5 + 1.7438/5}} = 2.6342$$

5 $t_{0.05/2,8} = 2.3060$ より，

$$T_0 = 1.7428 < t_{0.05/2,8} = 2.3060$$

したがって，H_0 は棄却されない．

(c) σ_1^2, σ_2^2 が未知

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/(n_1 - 1) + S_2^2/(n_2 - 1)}} \sim t(\phi),$$

$$\frac{1}{\phi} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1},\ \frac{1}{c} = 1 + \frac{(n_1 - 1)S_2^2}{(n_2 - 1)S_1^2}$$

2. **母分散の比 σ_1^2/σ_2^2 の検定**

$$\frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1),\ \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

より

$$F = \frac{\sigma_2^2 {S_1'}^2}{\sigma_1^2 {S_2'}^2} \sim F(n_1 - 1, n_2 - 1)$$

したがって，

1. 対立仮説が $H_1 : \sigma_1^2 \neq \sigma_2^2$ のときは，両側検定で，棄却域は

$$W = \{F : F > F_{n_2-1}^{n_1-1}(\frac{\alpha}{2})\} \cup \{F : F < F_{n_2-1}^{n_1-1}(1 - \frac{\alpha}{2})\}$$

2. 対立仮説が $H_1 : \sigma_1^2 > \sigma_2^2$ のとき棄却域は

$$W = \{F : F > F_{n_2-1}^{n_1-1}(\alpha)\}$$

3. 対立仮説が $H_1 : \sigma_1^2 < \sigma_2^2$ のとき棄却域は

$$W = \{F : F < F_{n_2-1}^{n_1-1}(1 - \alpha)\}$$

しかし，$F_{n_2-1}^{n_1-1}(1-\alpha)$ は数表にありません．そこで，この場合は，

$$F_{n_2-1}^{n_1-1}(1-\alpha) = \frac{1}{F_{n_1-1}^{n_2-1}(\alpha)}$$

を用いて計算します.

例題 4.4

A,B 2つの機械から製造された製品から，それぞれの大きさ 10，16 の標本を抽出して重量を調べたところ，分散がそれぞれ $5.23g^2$, $2.48g^2$ であった．母分散に有意差があるか．有意水準 5%で検定せよ.

解

$$\begin{aligned} n_A &= 10,\ S_A^2 = 5.23,\ S'^2 = \frac{10}{9}S_A^2 \\ n_B &= 16,\ S_B^2 = 2.24,\ S'^2 = \frac{16}{15}S_B^2 \end{aligned}$$

1　$H_0 : \sigma_1^2 = \sigma_2^2 H_1 : \sigma_1^2 \neq \sigma_2^2$

2　有意水準 $\alpha = 0.05$

3　統計量

$$F = \frac{\sigma_2^2 {S_1'}^2}{\sigma_1^2 {S_2'}^2} \sim F(n_1 - 1, n_2 - 1)$$

4　H_0 のもとで,

$$F_0 = \frac{\frac{10(5.23)}{9}}{\frac{16(2.24)}{15}} = 2.1967$$

5　$F_{15}^9(0.025) = 3.1227$ より,

$$F_0 = 2.1967 < F_{15}^9(0.05/2) = 3.1227$$

したがって，H_0 は棄却できない．すなわち両方の母分散に有意差はない.

統計学演習問題 9

1 環境学部の A，B で数学の試験をした．A クラスから 10 名，B クラスから 12 名の成績を無作為に選んだら次の表を得た．

A	71	79	92	91	87	79	77	89	71	84		
B	63	84	71	81	80	84	71	84	64	84	69	77

A，B の成績は，それぞれ正規分布 $N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$ に従っており，分散は同じであるとする．このとき，仮説

$$H_0 : \mu_1 = \mu_2$$

を有意水準 5% で両側検定をしよう．

2 環境学部の A，B で数学の試験をした．A クラスから 10 名，B クラスから 12 名の成績を無作為に選んだら次の表を得た．

A	71	79	92	91	87	79	77	89	71	84		
B	63	84	71	81	80	84	71	84	64	84	69	77

A，B の成績は，それぞれ正規分布 $N(\mu_1, {\sigma_1}^2), N(\mu_2, {\sigma_2}^2)$ に従っているとする．このとき，仮説

$$H_0 : \sigma_1 = \sigma_2$$

を有意水準 5% で左側検定をしよう．

4.3 比率の検定 (Test of proportions)

母比率の検定 (大標本の場合) 母集団の中で，ある属性に対して事象 A の起こる割合 p を事象 A の**母比率**といいます．この母比率に関する仮説を，標本値から検定することを考えます．

母比率が p の二項母集団から抽出された大きさ n の標本を $(X_1, \ldots, X_n)$ とします．ここで，

$$X_i = \begin{cases} 1 & A \text{ のとき} \\ 0 & \bar{A} \text{のとき} \end{cases}$$

とします．このとき，$X = X_1 + \cdots + X_n$ とすると，X は標本中 A であるものの個数を表す統計量で，$\dfrac{X}{n}$ は事象 A の**標本比率**といいます．そのとき，母比率 p について，$p_0 (0 \leq p_0 \leq 1)$ を既知の値として，帰無仮説

H_0: 「$p = p_0$」，対立仮説 H_1: 「$p \neq p_0$」

を検定することが問題となります．

母比率 p の二項母集団から大きさ n の標本 $(X_1, \ldots, X_n)$ をとり，$X = X_1 + \cdots + X_n$ とすると X は二項分布 $B(n, p)$ に従います．ここで n が十分大きいときにはラプラスの定理によって，X は近似的に正規分布 $N(np, np(1-p))$ に従い，標本比率 $\frac{X}{n} = \hat{p}$ は近似的に正規分布 $N\left(p, \frac{p(1-p)}{n}\right)$ に従います．よって，標準化を行うと

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

例題 4.5

サイコロを 600 回投げたところ，1 の目が 108 回出たという．1 の目が出る母比率 p は $\frac{1}{6}$ か有意水準 5%で検定せよ．

解答

1

H_0 : 「1 の目の出る確率 $p = \frac{1}{6}$」

H_1 : 「$p \neq \frac{1}{6}$」

2　有意水準 $\alpha = 0.05$

3　統計量

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

4　H_0 のもとで，$\hat{p} = \frac{108}{600} = 0.18$ より，

$$\begin{aligned} Z_0 &= \frac{0.18 - \frac{1}{6}}{\sqrt{\frac{\frac{1}{6}(1-\frac{1}{6})}{600}}} \\ &= 0.088 \end{aligned}$$

5　対立仮説より，標準正規分布の両側確率を用いる．

$$Z_0 < Z_{\frac{0.05}{2}} = 1.96$$

したがって，H_0 を容認.

母比率の差の検定

2 つの母集団 A, B の中で 1 つの特性 C を持つものの母比率を p_1, p_2 とする．この母集団からそれぞれ大きさ n_1, n_2 個の標本を抽出し，その特性を持つものの個数を X_1, X_2 とする．このとき，母比率について

帰無仮説　H_0: 「$p_1 = p_2$」と対立仮説　H_1: 「$p_1 \neq p_2$」

を検定することを考えます．帰無仮説のもとで，母比率の値 p_1, p_2 は未知ですが,

$$p_1 = p_2 = p,\ 1 - p = q$$

とおくと，n_1, n_2 がある程度大きければ X_1, X_2 の分布は正規分布 $(N(n_1p, n_1pq),\ N(n_2p, n_2pq)$ によってそれぞれ近似されるので，統計量

$$\bar{X_1} = \frac{X_1}{n_1} \sim (N(p, \frac{pq}{n_1})$$

$$\bar{X_2} = \frac{X_2}{n_2} \sim (N(p, \frac{pq}{n_2})$$

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \sim N(0, (\frac{1}{n_1} + \frac{1}{n_2})pq)$$

これより，

$$Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}} \sim N(0,1)$$

ただし，共通の母比率 p は未知数ですので，観測比率

$$p = \frac{X_1 + X_2}{n_1 + n_2}$$

を用います.

例題 4.6

テレビの視聴率調査で，ある番組について男性は 400 人中の無作為標本中 120 人が，女性は 500 人の無作為標本中 180 人が好きと答えた．実際に男女の好みに差があるといえるか，有意水準 5%で検定せよ．

解答 ある番組を男性が好きな比率を p_1，女性が好きな比率を p_2 とすると，$p_1 = \frac{120}{400}, p_2 = \frac{180}{500}$

1

H_0：「男女で好みに差がない」$p_1 = p_2$

H_1：「男女で好みに差がある」$p_1 \neq p_2$

2　有意水準 $\alpha = 0.05$

3　統計量

$$Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}} \sim N(0,1)$$

4　H_0 のもとで，$p = \frac{120+180}{400+500} = \frac{1}{3}$ より，

$$\begin{aligned} Z_0 &= \frac{120/400 - 180/500}{\sqrt{(\frac{1}{400} + \frac{1}{500})\frac{1}{3}(1-\frac{1}{3})}} \\ &= -1.897 \end{aligned}$$

5　対立仮説より，標準正規分布の両側確率を用いる．

$$|Z_0| < Z_{\frac{0.05}{2}} = 1.96$$

したがって，H_0 を容認.

統計学演習問題 10

1　ある政党の支持率は従来 28%であったが，最近の世論調査で無作為に抽出された 3,000 人有権者のうち支持率は 25%であった．支持率が低下したと判断すべきか，有意水準 5%で検定せよ．

2　あるテレビ番組の視聴率調査を男女別に行った．その結果，男性の無作為標本 200 人のうち 25 人が，女性の無作為標本 300 人のうち 20 人が見ていると答えた．このとき，男女の視聴率に差があるといえるか，有意水準 5%で検定せよ．

4.4　適合度検定 (Goodness of fit)

データにある確率分布をあてはめ，あてはまりのよさを検定するのが**適合度検定 (goodness of fit test)** です．この検定の問題に対して，標本は元のデータに対応します．また，想定した確率分布には，ある確率変数 X が対応しています．

(1) **多項分布に対する適合度の検定**

ある試行の結果，k 個の事象 $A_1, A_2, \ldots, A_k$ のいずれかが現われるとします．ここで，A_i が起こる確率を $P(A_i)$ とすると，

$$\begin{aligned} P(A_i) &= p_i \\ p_1 + p_2 + \cdots + p_k &= 1 \end{aligned}$$

となります．この試行を n 回独立に行なうとき，$A_1, A_2, \ldots, A_k$ がそれぞれ $n_1, n_2, \ldots, n_k$ 回現われる確率は

$$\binom{n}{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

ここで，$k = 2$ のときは，2 項分布に他なりません．この 2 項分布の一般化を**多項分布 (multinomial distribution)** といい，n 回の独立試行で事象 A_i が起こる回数を確率変数 X_i で表すと，

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

となります．

例題 4.7

それぞれの目が出る確率が等しいサイコロがある．これを 6 回投げたとき，1 から 6 までが 1 回ずつ現れる確率を求めよ．

解 各数字が現れる確率は $\frac{1}{6}$ で，1 から 6 までが 1 回ずつ現れる組み合わせは $\binom{6}{1,1,1,1,1,1}$ 通り．したがって，その確率は

$$\binom{6}{1,1,1,1,1,1} (\frac{1}{6})^6 = \frac{6!}{6^6} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 6 \cdot 6 \cdot 6 \cdot 6 \cdot 6} = \frac{5}{324}$$

次に，$A_1, A_2, \ldots, A_k$ の互いに排反な事象のいずれかが現われる多項分布を考えます．$P(A_i) = p_i$ とすると，大きさ n の標本のうち A_i に入る期待値は $np_i = m_i$ となります．一方，大きさ n の標本

のうち A_i の部分に入る個数を確率変数 X_i で表すと，次のことが知られています．$m \geq 5$ のとき，

$$\chi^2 = \frac{(X_1 - m_1)^2}{m_1} + \frac{(X_2 - m_2)^2}{m_2} + \cdots + \frac{(X_k - m_k)^2}{m_k}$$

は $\chi^2(k-1)$ に従う．

理論度数 m_i と実測度数 X_i がすべての i について近い値であれば，χ^2 は全体として小さな値となります．したがって，χ^2 が大きな値となったとき，その理論値 m_i に疑問が持たれます．このことから，次のような適合度の検定が得られます．

帰無仮説 $H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$
(p_{i0} は正数で $p_{10} + p_{20} + \cdots + p_{k0} = 1$ となる数)
対立仮説 $H_1 : p_1 = p_{11}, p_2 = p_{21}, \ldots, p_k = p_{k1}$
ただし，$(p_{11}, p_{21}, \ldots, p_{k1}) \neq (p_{10}, p_{20}, \ldots, p_{k0})$ である．

ここでは問題の性質上，片側検定にあたるものは考えられません．H_0 のもとで A_i に入る理論度数 m_i は，

$$m_i = np_{i0}$$

で与えられます．ここでは，n は十分大きな値で，すべての i に対して $m_i \geq 5$ となるとします．A_i に入る標本値が x_i であるとき
$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - np_{i0})^2}{np_{i0}} > \chi^2_{\alpha,k-1}$ ならば H_0 を棄却する．
これによって適合度が検定できます．

例題 4.8

あるサイコロを 600 回投げたところ，次のような表が得られた．各目の現れる確率が等しいと考えられるか，有意水準 0.05 で検定しよう．

目の数	1	2	3	4	5	6	計
回数	102	89	87	106	115	101	600

解 1
H_0 : 「各目の現れる確率は等しい」$(p_1, p_2, p_3, p_4, p_5, p_6 = \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
H_1 : 「各目の現れる確率は等しくない」$(p_1, p_2, p_3, p_4, p_5, p_6 \neq \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$

2 有意水準 $\alpha = 0.05$
3 統計量

$$\chi^2 = \sum_{i=1}^{6} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{6} \frac{X_i^2}{np_i} - n$$

4 H_0 のもとで，

$$\begin{aligned} \chi_0^2 &= \frac{102^2}{100} + \frac{89^2}{100} + \frac{87^2}{100} + \frac{106^2}{100} + \frac{115^2}{100} + \frac{101^2}{100} - 600 \\ &= 104.04 + 79.21 + 75.69 + 112.36 + 132.25 + 102.01 - 600 = 5.56 \end{aligned}$$

5　$\chi^2_{0.05,6-1} = 12.83$ より，

$$\chi^2_0 = 5.56 < \chi^2_{0.05,5} = 11.07$$

したがって，H_0 を容認.

統計学演習問題 11

1 ある遺伝形質は $A:B:C:D=9:3:3:1$ のメンデル比に従って現われるとされているが，実験の結果次の表を得た．メンデル比に従っているといえるか，有意水準 5%で検定しよう．

遺伝形質	A	B	C	D	計
観測度数	243	72	78	15	408

(2) **確率分布に対する適合度の検定**

ここでは，ある分布が正規分布に従う，あるいはポワソン分布に従う，ということ自体が帰無仮説となる適合度検定を考えます．つまり，

帰無仮説　H_0：「ある分布 D に従う」

を設定します．D の分布は既知であって，母数 $\theta_1,\theta_2,\ldots,\theta_i$ を含んでいるとします．例えば，正規分布では μ,σ^2 の 2 個の母数を含み，これらの値は不明であるとします．

次に排反な各階級 $A_1,A_2,\ldots,A_k$ に入る個数 $(X_1,X_2,\ldots,X_k)$ の実現値を $(x_1,x_2,\ldots,x_k)$ とし，母数 θ_i をこの値を用いて推定します．つまり，

$$\theta_i=\hat{\theta}_i(x_1,x_2,\ldots,x_k)\quad(i=1,2,\ldots,l)$$

この θ_i を用いて各階級 $A_1,A_2,\ldots,A_k$ に入るべき期待度数 $m_1,m_2,\ldots,m_k$ を求めます．ここで，$m_i=np_{i0}$. つまり，

$$\chi^2=\sum_{i=1}^{k}\frac{(X_i-m_i)^2}{m_i}$$

を求めます．このとき，$\chi^2\sim\chi^2_{k-l-1}$ であることが分かっています．そして，これを用いて H_0 の検定を行ないます．

例題 4.9

ある軍隊の 10 個の部隊において，1 年間に馬に蹴られて死亡した兵士の数とその部隊数を 10 年間調べた結果次のような表になった．

死亡者数	0	1	2	3	4	計
部隊数	109	65	22	3	1	200

この表はポワソン分布に従うか，有意水準 5%で検定しよう．

1

H_0：「ポワソン分布 $P(\lambda)$ に従っている」

2　有意水準 $\alpha=0.05$

3　統計量

この表をポワソン分布とみて，死亡数の理論値を求める．これがポワソン分布 $P(\lambda)$ によるものと考えて，λ の値を推定する．死亡者数 k のときの確率を p_k とすると，

$$\sum_{k=0}^{\infty}kp_k=E(X)=\lambda$$

死亡者数	k	0	1	2	3	4	計
部隊数	f_k	109	65	22	3	1	200
	kf_k	0	65	44	9	4	122
	p_k	0.5435	0.3313	0.1011	0.0206	0.0031	
理論度数	m_k	108.7	66.3	20.2	4.1	0.6	

ここで，$np_k \approx f_k$ より $\sum_k kf_k \approx \lambda n$．これより平均値 λ は

$$\lambda \approx \frac{1}{n}\sum_k kf_k = \frac{122}{200} = 0.61$$

死亡者数	k	0	1	2	3	4	計
部隊数	x_k	109	65	22	3	1	200
理論度数	m_k	108.7	66.3	20.2	4.1	0.6	

この表で，$k \geq 3$ の所の m_k は単独で 5 よりも小さいので，χ^2 検定ができない．そこで，右から順に m_i を加えて 5 を越すまで合併すると，$k \geq 2$ の階級を 1 つにしなければならない．したがって，

$$\chi^2 = \sum_{i=0}^{2} \frac{(x_i - m_i)^2}{m_i}$$

4　H_0 のもとで，

$$\begin{aligned} \chi_0^2 &= \frac{(109-108.7)^2}{108.7} + \frac{(65-66.3)^2}{66.3} + \frac{(26-25)^2}{25} \\ &= 0.066 \end{aligned}$$

5　$\chi^2_{0.05,3-1-1} = 3.84$ より，

$$\chi_0^2 = 0.066 < \chi^2_{0.05,1} = 3.84$$

したがって，H_0 を容認．

母数 λ が標本から 1 個推定されたので，自由度は $3-1-1=1$ となる．

(3) 独立性の検定

母集団の要素は，すべて A, B の 2 種類の属性をもち，A, B はそれぞれ排反な $A_1, \ldots, A_k$ および $B_1, \ldots, B_l$ に分かれているとします．母集団から大きさ n の標本を抽出して，$A_i \cap B_j$ に入る観測度数を x_{ij} とすると，次の表のように行列の形に整理できる．

	B_1	B_2	$\cdots$	B_l	和
A_1	x_{11}	x_{12}	$\cdots$	x_{1l}	x_1
A_2	x_{21}	x_{22}	$\cdots$	x_{2l}	x_2
A_3	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	x_{k1}	x_{k2}	$\cdots$	x_{kl}	x_k

ここで，$x_{i.}, x_{.j}$ は周辺度数である．このような表を $k \times l$ **分割表 (contingency table)** という．

これを用いて，母集団の属性 A と B が無関係であるかを調べることを**独立性の検定**という．独立性の検定には適合度の検定を応用することができる．

対 A_i, B_j の出現度数の確率変数を X_{ij}，A_i, B_j の実現する確率を p_i, q_j．また，A_i, B_j が同時に起こる確率を P_{ij} とする．

ここで，次のような適合度の検定を考える．

帰無仮説　：　「属性 A, B は独立である」

対立仮説　：　「属性 A, B は従属である」

帰無仮説 H_0 のもとで

$$P_{ij} = P_r(A_i \cap B_j) = P_r(A_i)P_r(B_j) = p_i q_j$$

が成り立つ．ここで，p_i, q_j は母数なのでこれを最尤法によって推定すると，それらの推定値は

$$\hat{p_i} = \frac{x_{i.}}{n},\ \hat{q_j} = \frac{x_{.j}}{n}$$

で与えられる．このとき，n が十分大きければ，帰無仮説 H_0 のもとで統計量

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(X_{ij} - nP_{ij})^2}{nP_{ij}}$$

が自由度 $(k-1)(l-1)$ のカイ 2 乗分布に従うことが知られている．観測度数 x_{ij} を用いると，統計量 χ^2 の実現値は

$$\begin{aligned}
\chi_0^2 &= \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(x_{ij} - n\hat{p_i}\hat{q_j})^2}{n\hat{p_i}\hat{q_j}} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{l} \left\{ \frac{x_{ij}^2}{n\hat{p_i}\hat{q_j}} - 2x_{ij} + n\hat{p_i}\hat{q_j} \right\} = n\left\{ \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{x_{ij}^2}{x_{i.}x_{.j}} - 1 \right\}
\end{aligned}$$

となる．

統計学演習問題 12

1　ある軍隊の 10 個の部隊において，1 年間に馬に蹴られて死亡した兵士の数とその部隊数を 10 年間調べた結果次のような表になった．

死亡者数	0	1	2	3	4	計
部隊数	142	99	46	11	3	300

この表はポワソン分布に従うか，有意水準 5%で検定しよう．

2　350 人の大人を無作為に抽出して，飲酒と喫煙について答えてもらった．その際，飲酒の程度を低い方から A_1, A_2, A_3 と 3 段階に分け，喫煙の程度は低い方から B_1, B_2, B_3, B_4 と 4 段階に分けた．結果は次の通りであった．飲酒と喫煙は関係があるか，有意水準 5%で検定しよう．

	B_1	B_2	B_3	B_4	計
A_1	39	54	49	17	159
A_2	27	43	40	9	119
A_3	14	23	15	20	72
計	80	120	104	46	350

4.5　検定に用いる統計量 (Test statistic)

母平均の検定

母集団が正規分布 $N(\mu, \sigma^2)$ に従っていて，母分散 σ^2 が既知の場合

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

母集団が正規分布に従っていて $N(\mu, \sigma^2)$，母分散 σ^2 が未知の場合．母分散 σ^2 の不偏推定量 S'^2 を用いる

$$T = \frac{\bar{X} - \mu}{\sqrt{S'^2/n}} \sim t(n-1)$$

母分散の検定

母集団が正規分布 $N(\mu, \sigma^2)$ に従っていて，母平均 μ が既知の場合

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi^2(n)$$

母集団が正規分布 $N(\mu, \sigma^2)$ に従っていて，母平均 μ が未知の場合

$$\chi^2 = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

母平均の差の検定

2つの母集団がそれぞれ正規分布 $N(\mu_1,\sigma_1^2)$,$N(\mu_2,\sigma_2^2)$ に従っていて，母分散 σ_1^2,σ_2 が既知の場合

$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim N(0,1)$$

2つの母集団がそれぞれ正規分布 $N(\mu_1,\sigma_1^2)$,$N(\mu_2,\sigma_2^2)$ に従っていて，母分散 σ_1^2,σ_2 が未知であるが等しいとみなせる場合

$$T=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\hat{\sigma}^2/n_1+\hat{\sigma}^2/n_2}}\sim t(n_1+n_2-2),\ \hat{\sigma}^2=\frac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}.$$

母分散の比の検定

2つの母集団がそれぞれ正規分布 $N(\mu_1,\sigma_1^2)$,$N(\mu_2,\sigma_2^2)$ に従っている場合

$$F=\frac{\sigma_2^2{S_1'}^2}{\sigma_1^2{S_2'}^2}\sim F(n_1-1,n_2-1)$$

母比率の検定

二項母集団からの大きさ n の標本のうち属性 A を持つものの個数を確率変数 X で表すと，$X\sim B(n,p)$ に従う．n が十分大きいときは X は近似的に正規分布 $N(np,np(1-p))$ に従う (ラプラスの定理). これを用いて，母比率 p の検定を行う．

$$Z=\frac{X-np}{\sqrt{np(1-p)}}\sim N(0,1)$$

ただし，$p=\frac{X_1+X_2}{n_1+n_2}$

母比率の差の検定

2つの母集団 A,B の中で 1 つの特性 C を持つものの母比率を p_1,p_2 とする．この母集団からそれぞれ大きさ n_1,n_2 個の標本を抽出し，その特性を持つものの個数を X_1,X_2 とする．

$$Z=\frac{X_1/n_1-X_2/n_2}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})p(1-p)}}\sim N(0,1)$$

多項分布に対する適合度の検定

$A_1,A_2,\ldots,A_k$ の互いに排反な事象のいずれかが現れる多項分布において，$P(A_i)=p_i$ とすると，大きさ n の標本のうち A_i に入る期待値は $np_i=m_i$ である．また，大きさ n の標本のうち A_i に入る個数を確率変数 X_i で表すと，$m\geq 5$ のとき，

$$\chi^2=\sum_{i=1}^{k}\frac{(X_i-np_i)^2}{np_i}=\sum_{i=1}^{k}\frac{X_i^2}{np_i}-n\sim\chi^2(k-1)$$

確率分布に対する適合度の検定

帰無仮説　H_0：「ある分布 D に従う」において，D の分布の型は既知であって，母数 $\theta_1,\ldots,\theta_l$ を含んでいるとする．次に，$A_1,A_2,\ldots,A_k$ の互いに排反な事象のいずれかが現れる個数 $(X_1,\ldots,X_k)$ の実現値を $(x_1,\ldots,x_k)$ とし，母数 θ_i を推測する．そして，この θ_i を用いて A_i に入るべき期待度数 $m_1,\ldots,m_k$ を計算する．このとき，

$$\chi^2=\sum_{i=1}^{k}\frac{(X_i-m_i)^2}{np_i}\sim\chi^2(k-l-1)$$

独立性の検定

対 A_i, B_j の出現度数の確率変数を X_{ij}，A_i, B_j の実現する確率を p_i, q_j．また，A_i, B_j が同時に起こる確率を P_{ij} とする．$nP_{ij} \geq 5$ のとき，

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{X_{ij} - nP_{ij}}{nP_{ij}} \sim \chi^2((k-1)(l-1)), \quad P_{ij} = p_i q_j$$

第5章　演習問題解答

1 スタージスの式から

階級数 $= 1 + \frac{\log 100}{\log 2} = 1 + 6.64 = 7.64$ また最大値 440 最小値 300 より，階級幅は

階級幅 $= \frac{440-300}{7.64} = 18.4$ となるので，階級幅を 18 ととることにします．これより度数分布表を作成します.

表 5.1: 度数分布表

階級	階級値	度数	相対度数	累積度数	累積相対度数
300 ∼ 318	309	2	0.02	2	0.02
318 ∼ 336	327	10	0.1	12	0.12
336 ∼ 354	345	25	0.25	37	0.37
359 ∼ 372	363	31	0.31	68	0.68
372 ∼ 390	381	8	0.08	76	0.76
390 ∼ 408	399	18	0.18	94	0.94
408 ∼ 426	417	5	0.85	99	0.99
426 ∼ 444	435	1	0.01	100	1.00

平均値

$$\begin{aligned} \overline{x} &= \frac{1}{100}[318 \cdot 2 + 336 \cdot 10 + 354 \cdot 25 + 363 \cdot 31 + 381 \cdot 8 + 399 \cdot 18 + 417 \cdot 5 + 435 \cdot 1] \\ &= 365.16 \end{aligned}$$

最大値 440

最小値 300

中央値 $\frac{360 + 360}{2} = 360$

最頻値 363

図 5.1: ヒストグラム

図 5.2: 累積度数分布表

問題解答 2

2.

電卓を使う場合は必ず途中の値を書く必要があります．また，計算は小数点以下 2 桁までで表わすことにします．

$T_x = 841 \quad T_y = 806$

$\overline{x} = 35.04 \quad \overline{y} = 33.58$

$T_{xx} = \sum_{i=1}^{24} x_i^2 = 45553 \quad T_{yy} \sum_{i=1}^{24} y_i^2 = 36990$

$s_x = \sqrt{\frac{1}{24} T_{xx} - (\overline{x})^2} = 25.88 \quad s_y = \sqrt{\frac{1}{24} T_{yy} - (\overline{y})^2} = 20.34$

$T_{xy} = \sum_{i=1}^{24} x_i y_i = 37192$

これより

$$\begin{aligned} s_{xy} &= \frac{1}{n} T_{xy} - \frac{T_x}{n} \frac{T_y}{n} \\ &= \frac{1}{24} \cdot 37192 - \frac{841}{24} \cdot \frac{806}{24} = 372.85 \end{aligned}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{372.85}{25.88 \cdot 20.34} = 0.71$$

これより正の相関でかなり強い相関があるといえる.

問題解答 3

3

階級数 $= 1 + \frac{\log n}{\log 2} = 1 + \frac{\log 24}{\log 2} = 1 + 4.58 = 5.58$
また x の最大値 109 最小値 13 より
階級幅 $= \frac{109-13}{5.58} = 17.20$
これより x の階級幅を 17 と取ります. また y の最大値 65 最小値 5 より

$$\text{階級幅} = \frac{65-5}{5.58} = 10.75$$

これより y の階級幅を 10 と取ります.

表 5.2: 相関表

	x	$10 \sim 27$	$27 \sim 44$	$44 \sim 61$	$61 \sim 78$	$78 \sim 95$	$95 \sim 112$	
y	階級値	18.5	35.5	52.5	69.5	86.5	103.5	計
$0 \sim 10$	5	3						3
$10 \sim 20$	15	6						6
$20 \sim 30$	25	1	2					3
$30 \sim 40$	35			1				1
$40 \sim 50$	45	1	1					2
$50 \sim 60$	55	2	3			2		7
$60 \sim 70$	65					1	1	2
		13	7	1		3	1	24

$T_x = 841 \quad T_y = 806$
$\overline{x} = 35.04 \quad \overline{y} = 33.58$
$T_{xx} = \sum_{i=1}^{24} x_i^2 = 45553 \quad T_{yy} \sum_{i=1}^{24} y_i^2 = 36990$
$s_x = \sqrt{\frac{1}{24} T_{xx} - (\overline{x})^2} = 25.88 \quad s_y = \sqrt{\frac{1}{24} T_{yy} - (\overline{y})^2} = 20.34$
$T_{xy} = \sum_{i=1}^{24} x_i y_i = 37192$

$$\begin{aligned} s_{xy} &= \frac{1}{n} T_{xy} - \frac{T_x}{n} \frac{T_y}{n} \\ &= \frac{1}{24} \cdot 37192 - \frac{841}{24} \cdot \frac{806}{24} = 372.85 \end{aligned}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{372.85}{25.88 \cdot 20.34} = 0.71$$

これより，x 上の y の回帰直線は

$$y - 33.58 = \frac{372.85}{670.24}(x - 35.04) = 0.56(x - 35.04)$$

したがって

$$y = 0.56x + 13.96$$

問題解答 4

1.

(a) 標準化する．

$$\begin{aligned} P_r(X \leq 90) &= P_r\left(\frac{X-80}{6} \leq \frac{90-80}{6}\right) \\ &= P_r(Z \leq 1.67) = P_r(-\infty < Z < 0) + P_r(0 \leq Z \leq 1.67) = 0.95 \end{aligned}$$

(b)

$$\begin{aligned} P_r(|X-80| \leq 12) &= P_r\left(\frac{|X-80|}{6} \leq \frac{12}{6}\right) \\ &= P_r(|Z| \leq 2) = 2P_r(0 \leq Z \leq 2) = 2(0.477) = 0.95 \end{aligned}$$

(2) X を一人一日当たりの水需要量とすると， $X \sim N(210, 21^2)$．これより一人当たりの水需要量 (夏期を除く) が 250(l/人) 以上になる確率は $P_r(X \geq 250)$ で与えられる．したがって，

$$\begin{aligned} P_r(X \geq 250) &= P_r\left(\frac{X-210}{21} \geq \frac{250-210}{21}\right) \\ &= P_r(Z \geq 1.90) = \frac{1}{2} - P_r(0 < Z < 1.96) = 0.5 - 0.475 = 0.025 \end{aligned}$$

問題解答 5

1.

$$\begin{aligned} \bar{X} &= \frac{1}{10}(110 + 121 + 133 + 124 + 126 + 118 + 112 + 125 + 131 + 120) = 122(\text{cm}) \\ U^2 &= \frac{1}{9}\left((110-122)^2 + (121-122)^2 + \cdots + 120 - 122)^2\right) \\ &= 55.111(\text{cm}) \end{aligned}$$

また，$S^2 = \frac{9}{10}U^2 = 49.5999$ より，$S = 7.043$ となります．

問題解答 6

1 ある水域の一定区間における水質 BOD を X とおくと，$X \sim N(\mu, 6.25)$．又，標本数は 15 で，標本平均 $\overline{X} = 7.2$ より，

$$\overline{X} \sim N(\mu, \frac{6.25}{15})$$

ここで，$\overline{X}$ を標準化すると，

$$Z = \frac{X-\mu}{\sqrt{\frac{6.25}{15}}} = \frac{7.25-\mu}{\sqrt{\frac{6.25}{15}}}$$

これより，

$$P_r(|Z| \leq z_{\frac{\alpha}{2}}) = 0.95$$

を満たす，$z_{\frac{\alpha}{2}}$ を標準正規分布表を用いて求めると，$z_{\frac{\alpha}{2}} = 1.96$ したがって，95%信頼区間は

$$|Z| = |\frac{7.25-\mu}{\sqrt{\frac{6.25}{15}}}| \leq 1.96$$

つまり

$$7.25 - 1.96\sqrt{\frac{6.25}{15}} \leq \mu \leq 7.25 + \sqrt{\frac{6.25}{15}}$$

2　標準偏差が 2.5 より，母分散 $\sigma^2 = 6.25$ は既知である．この母集団から無作為に選んだ標本 X_i は $X_i \sim N(\mu, 6.25)$ の正規分布に従っていると考えることができる．したがって，

$$\bar{X} \sim N(\mu, \sigma^2/5)$$

となる．ここで，$\bar{X}$ を求めると，

$$\bar{X} = \frac{1}{4}[145.3 + 145.1 + 145.4 + 146.2] = \frac{582}{4} = 145.5$$

標準化を行なうと，

$$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/5}} \sim N(0,1)$$

となる．95%信頼区間より，$P_r(|Z| \leq z_{\frac{\alpha}{2}}) = 0.95$. また，$z_{\frac{0.05}{2}} = 1.96$. したがって，

$$\bar{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{4}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{4}}$$

$$145.5 - 1.96\sqrt{6.25/4} \leq \mu \leq 145.5 + 1.96\sqrt{6.25/4}$$

$$143.05 \leq \mu \leq 147.95$$

3　母平均 $\mu = 146$ であるが母分散 σ^2 は未知である．この母集団から無作為に選んだ標本 X_i は $X_i \sim N(146, \sigma^2)$ の正規分布に従っていると考えることができる．したがって，

$$\bar{X} \sim N(146, \sigma^2/4)$$

となる．ここで，$\bar{X}$ を求めると，

$$\bar{X} = \frac{1}{4}[145.3 + 145.1 + 145.4 + 146.2] = \frac{582}{4} = 145.5$$

母分散の推定に S'^2 を用いると，

$$T = \frac{\bar{X}-\mu}{\sqrt{S'^2/4}} \sim t_{n-1,\alpha/2}$$

となる．そこで，S'^2 を求めると，

$$\begin{aligned} S'^2 &= \frac{1}{3}[(145.3-145.5)^2+(145.1-145.5)^2+(145.4-145.5)^2+(146.2-145.5)^2] \\ &= \frac{1}{3}(0.04+0.16+0.01+0.49)=0.23 \end{aligned}$$

となる．

95%信頼区間より，$P_r(|T| \leq t_{n-1,\alpha/2}) = 0.95$. また，$t_{3,0.005/2} = 3.18$. したがって，

$$\bar{X} - t_{3,0.05/2}\sqrt{\frac{S'^2}{4}} \leq \mu \leq \bar{X} + t_{3,0.05/2}\sqrt{\frac{S'^2}{4}}$$

$$145.5 - 3.18\sqrt{0.23/4} \leq \mu \leq 145.5 + 3.18\sqrt{0.23/4}$$

$$144.73 \leq \mu \leq 146.26$$

問題解答 7

1 標本比率は $\hat{p} = \frac{180}{900} = 0.2$．また，$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96$ であるから，十分大きな n に対して，統計量 $\bar{X}$ の分布が正規分布 $N(p, \frac{pq}{n})$ で近似される．したがって，与えられた α に対して

$$P(\frac{|\bar{X}-p|}{\sqrt{pq}n} \leq z_{\frac{\alpha}{2}}) = 1-\alpha$$

が成り立つ．これより，

$$\bar{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

が成り立つ．ここで，$\bar{X}$ と p を $\bar{p}$ で置き換えると，

$$(\bar{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

これより，

$$\left(0.2 - 1.96\sqrt{\frac{(0.2)(1-0.2)}{900}},\ 0.2 + 1.96\sqrt{\frac{(0.2)(1-0.2)}{900}}\right)$$

より，$(0.174, 0.226)$ となる．

1 標本比率は $\hat{p} = \frac{187}{300} = 0.623$．また，$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96$ であるから，十分大きな n に対して，統計量 $\bar{X}$ の分布が正規分布 $N(p, \frac{pq}{n})$ で近似される．したがって，与えられた α に対して

$$P(\frac{|\bar{X}-p|}{\sqrt{pq}n} \leq z_{\frac{\alpha}{2}}) = 1-\alpha$$

が成り立つ．これより，

$$\bar{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

が成り立つ．ここで，$\bar{X}$ と p を $\bar{p}$ で置き換えると，

$$(\bar{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

これより，

$$\left(0.623 - 1.96\sqrt{\frac{(0.623)(1-0.623)}{300}},\ 0.623 + 1.96\sqrt{\frac{(0.623)(1-0.623)}{300}}\right)$$

より，$(0.568, 0.678)$ となる．

問題解答 8

1 この工場の製品を X とすると，$X \sim N(\mu, 0.20^2)$ であることが分かる．この工場から 16 個の製品を取り出したとき，$\bar{X}$ をそれらの標本平均とすると，

$$\bar{X} = 7.09,\ \bar{X} \sim N(\mu, 0.20^2/16)$$

となる．次に，この日の製品が規格から外れているかの検定を行なう．

1

$$\begin{aligned} H_0 &: \mu = 7 \\ H_1 &: \mu \neq 7 \end{aligned}$$

2　有意水準 $\alpha = 0.05$

3　統計量 σ^2 が既知より，

$$Z = \frac{\bar{X} - \mu}{\sigma^2/n} \sim N(0, 1)$$

4　H_0 のもとで，

$$Z_0 = \frac{7.09 - 7.00}{\sqrt{0.20^2/16}} = \frac{4(0.09)}{0.2} = 1.8$$

5

図 5.3: 正規分布

$z_{\frac{0.05}{2}} = 1.96$ より，H_0 を容認する．
95%信頼区間は

$$7.09 - 1.96\sqrt{\frac{0.20^2}{16}} \leq \mu \leq 7.09 - 1.96\sqrt{\frac{0.20^2}{16}}$$

より，

$$6.99 \leq \mu \leq 7.19$$

2　この工場の製品を X とすると，$X \sim N(\mu, 0.0016)$ であることが分かる．この工場から 8 個の製品を取り出したとき，$\bar{X}$ をそれらの標本平均とすると，

$$\bar{X} = \frac{1}{8}(11.97 + 12.02 + \cdots + 12.05) = 12.028$$

となる．次に，この日の製品の直径の分散は 0.0016 より大きいといえるかの検定を行なう．
1

$$\begin{aligned} H_0 &: \sigma^2 = 0.0016 \\ H_1 &: \sigma^2 > 0.0016 \end{aligned}$$

2　有意水準 $\alpha = 0.05$
3　統計量 μ が既知で σ^2 の検定を行なうので

$$\chi^2 = \frac{nS^2}{\sigma^2} \sim \chi^2_{\alpha, n-1}$$

4　H_0 のもとで

$$S^2 = \frac{1}{8}[(11.97 - 12.02)^2 + \cdots + (12.05 - 12.028)^2] = 0.0021$$

$$\chi_0^2 = \frac{8(0.0021)}{0.0016} = 10.5$$

5

$$\chi_0^2 > \chi^2_{0.05,7} = 14.07$$

となり，H_0 は棄却される．

図 5.4: χ 2 乗分布

95%信頼区間は

$$\chi^2_{1-0.05/2,8-1} \leq \frac{0.0168}{\sigma^2} \leq \chi_{0.05/2,8-1}$$

より,

$$1.690 \leq \frac{0.0168}{\sigma^2} \leq 16.01$$

$$0.0010 \leq \sigma^2 \leq 0.0099$$

問題解答 9

1

$$\begin{aligned} n_A &= 10, \bar{X} = 82, S_A^2 = 54.41 \\ n_B &= 12, \bar{Y} = 76, S_B^2 = 59.17 \end{aligned}$$

母平均の差の検定である．$\sigma_1^2 = \sigma_2^2$ より,

1 $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

2 有意水準 $\alpha = 0.05$

3 統計量

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}^2}{n_A} + \frac{\hat{\sigma}^2}{n_B}}} \sim t_{n_A+n_B-2}$$

ただし,

$$\hat{\sigma}^2 = \frac{n_A S_A^2 + n_B S_B^2}{n_A + n_B - 2}$$

4　H_0 のもとで,

$$T_0 = \frac{82 - 76}{\sqrt{62.7/10 + 62.7/12}} = \frac{6}{\sqrt{6.27 + 5.23}} = 1.77$$

5　$t_{0.05/2,20} = 2.23$ より,

$$T_0 = 1.77 < t_{0.05/2,20} = 2.09$$

したがって，H_0 は棄却されない.

2

$$\begin{aligned} n_A &= 10, \bar{X} = 82, S_A^2 = 54.41, {S'_A}^2 = 60.44 \\ n_B &= 12, \bar{Y} = 76, S_B^2 = 59.17, {S'_B}^2 = 64.53 \end{aligned}$$

母分散の比の左側検定である．$\sigma_1^2 < \sigma_2^2$

1　$H_0 : \sigma_A^2 = \sigma_B^2$
　　$H_1 : \sigma_A^2 < \sigma_B^2$

2　有意水準 $\alpha = 0.05$

3　統計量

$$F = \frac{\sigma_B^2 {S'_A}^2}{\sigma_A^2 {S'_B}^2} \sim F_{n_A - 1, n_B - 1}$$

4　H_0 のもとで,

$$F_0 = \frac{60.44}{64.53} = 0.936$$

5

$$F_{1-0.05,10-1,12-1} = \frac{1}{F_{0.05,12-1,10-1}} = \frac{1}{3.105} = 0.322$$

より,

$$F_0 = 0.936 > F_{1-0.05,10-1,12-1} = 0.322$$

したがって，H_0 は棄却されない.

一般に,

$$F_{\alpha,n_1,n_2} = \frac{1}{F_{1-\alpha,n_2,n_1}}$$

が成り立つ.

問題解答 10

1

1

現在の政党支持率を p とする.

H_0 :「政党支持率は変わらない」$p = 0.28$

H_1 :「政党支持率は変わった」　$p \neq 0.28$

2 有意水準 $\alpha = 0.05$

3 統計量　標本数が 3000 と大標本の場合と考えられるので，標本比率 $\hat{p}$ は近似的に正規分布 $N(p, \frac{p(1-p)}{n})$ に従う．よって，標準化を行うと

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

4 H_0 のもとで，$\hat{p} = 0.25$ より，

$$Z_0 = \frac{0.25 - 0.28}{\sqrt{\frac{0.28(1-0.28)}{3000}}} = -3.65$$

5

$$|Z_0| > Z_{\frac{0.05}{2}} = 1.96$$

したがって，H_0 を容認できない．

2

1

ある番組を好きな男性の割合を p_1，女性が好きな割合を p_2 とする．H_0：「視聴率に男女差はない」 $p_1 = p_2$

H_1：「視聴率に男女差はある」 $p_1 \neq p_2$

2 有意水準 $\alpha = 0.05$

3 母比率の差の検定より，

$$Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}} \sim N(0,1)$$

4 H_0 のもとで，$p = \frac{25+20}{200+300} = 0.09$. よって

$$Z_0 = \frac{25/200 - 20/300}{\sqrt{(\frac{1}{25} + \frac{1}{20})(0.09)(0.91)}} = 0.005$$

5

$$Z_0 < z_{\frac{0.05}{2}} = 1.96$$

したがって，H_0 の視聴率に男女差はないを容認

問題解答 11

1 遺伝形質　$A : B : C : D = 9 : 3 : 3 : 1$

1

H_0：「メンデル比に従っている」$(p_1, p_2, p_3, p_4 = \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$

H_1：「メンデル比に従っていない」$(p_1, p_2, p_3, p_4 \neq \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$

2　有意水準 $\alpha = 0.05$

3　統計量

$$\chi^2 = \sum_{i=1}^{4} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{4} \frac{X^2}{np_i} - n$$

4　H_0 のもとで，

$$\begin{aligned}\chi_0^2 &= \frac{243^2}{229.5}+\frac{72^2}{76.5}+\frac{78^2}{76.5}+\frac{15^2}{25.5}-408\\ &= 257.294+67.764+79.529+8.824-408=5.411\end{aligned}$$

5　$\chi^2_{0.05,4-1}=7.81$ より，

$$\chi_0^2=5.411<\chi^2_{0.05,3}=7.81$$

したがって，H_0 を容認.

問題解答 12

1

1

H_0：「ポワソン分布 $P(\lambda)$ に従っている」

2　有意水準 $\alpha=0.05$

3　統計量

この表をポワソン分布とみて，死亡数の理論値を求める．これがポワソン分布 $P(\lambda)$ によるものと考えて，λ の値を推定する．死亡者数 k のときの確率を p_k とすると，

$$\sum_{k=0}^{\infty} kp_k = E(X)=\lambda$$

死亡者数	k	0	1	2	3	4	計
部隊数	f_k	142	99	46	11	2	300
	kf_k	0	99	92	33	8	232
	p_k	0.473	0.33	0.153	0.036	0.0066	
理論度数	m_k	141.9	99	45.9	10.8	1.98	

ここで，$np_k\approx f_k$ より $\sum_k kf_k\approx\lambda n$．これより平均値 λ は

$$\lambda\approx\frac{1}{n}\sum_k kf_k=\frac{232}{300}=0.77$$

死亡者数	k	0	1	2	3	4	計
部隊数	f_k	142	99	46	11	2	300
理論度数	m_k	141.9	99	45.9	10.8	1.98	

この表で，$k=4$ の所の m_k は単独で 5 よりも小さいので，χ^2 検定ができない．そこで，右から順に m_i を加えて 5 を越すまで合併すると，$k\geq 3$ の階級を 1 つにしなければならない．したがって，

$$\chi^2=\sum_{i=0}^{3}\frac{(x_i-m_i)^2}{m_i}$$

4　H_0 のもとで,

$$\begin{aligned}\chi_0^2 &= \frac{(142-141.9)^2}{141.9}+\frac{(99-99)^2}{99}+\frac{(46-45.9)^2}{45.9}+\frac{(13-12,78)^2}{12.78}\\ &= 0.0040\end{aligned}$$

5　$\chi^2_{0.05,3-1-1}=3.84$ より,

$$\chi_0^2 = 0.0040 < \chi^2_{0.05,1} = 3.84$$

したがって，H_0 を容認.

母数 λ が標本から 1 個推定されたので，自由度は $3-1-1=1$ となる.

2

1

H_0 :「飲酒と喫煙とは独立である.」

H_1 :「飲酒と禁煙とは独立ではない.」

2　有意水準 $\alpha = 0.05$

3　統計量

$$\chi^2 = \sum_{i=1}^{350}\frac{(X_{ij}-nP_{ij})^2}{nP_{ij}}$$

4　H_0 のもとで,

$$\begin{aligned}\chi_0^2 &= 350[\frac{1}{159}(\frac{39^2}{80}+\frac{54^2}{120}+\frac{49^2}{104}+\frac{17^2}{46})\\ &+ \frac{1}{119}(\frac{27^2}{80}+\frac{43^2}{120}+\frac{40^2}{104}+\frac{9^2}{46})+\frac{1}{72}(\frac{14^2}{80}+\frac{23^2}{120}+\frac{15^2}{104}+\frac{20^2}{46})]-1\\ &= 367.55\end{aligned}$$

4　自由度 $(3-1)\times(4-1)=6$ より，$\chi^2_{0.05,6}=12.59$.

5　$\chi_0^2 = 367.55 > \chi^2_{0.05,6} = 12.59$ より，H_0 は棄却される．したがって，飲酒と喫煙には関係がある.